ABSTRACT

Title of dissertation: LINK BETWEEN DYNAMICS AND FUNCTION IN SINGLE AND MULTI-SUBUNIT ENZYMES

Jie Chen, Doctor of Philosophy, 2010

Dissertation directed by: Professor Devarajan Thirumalai
Biophysics Program
Institute for Physical Science and Technology

Biopolymers, such as proteins and DNA, are polymers whose three-dimensional conformations define their biological functions. Current emphasis on structures has greatly advanced our understanding of the functions of biopolymers. However, there is a need to understand the deeper link between biopolymer dynamics and function, because in water and under cellular conditions, everything that biopolymers do can be understood in terms of "the jigglings and wigglings of atoms". These motions arise from thermal noise in the solvent and due to intrinsic motion of the enzymes. In biological systems, the motions are often highly regulated to ensure that cellular processes are executed over the required time scales. For enzymes, which are essentially proteins that catalyze chemical reactions or generate mechanical work, conformational fluctuations are coupled at various stages through interactions with ligands during the catalytic cycle. We have studied two different enzymes, dihydrofolate reductase (DHFR), which catalyzes reduction of dihydrofolate to tetrahydrofolate, and RNA polymerase (RNAP from bacteria and Pol II from yeast), which is respon-

sible for RNA synthesis using DNA as a template. In order to study the link between dynamics and function we have developed new methods and extended a variety of computational techniques. For DHFR, we use both evolutionary imprints (SCA) and structure-based perturbation method (SPM) to extract a network of residues that facilitate the transitions between two distinct conformational states (closed and occluded states). The transition kinetics and pathways connecting the closed and occluded states are described using Brownian dynamics (BD) simulation. We found the sliding motion of Met20 loop across helix 2 is involved in the forward and reverse transitions between the closed and occluded states. We also found that cross-linking M16-G121 inhibits both the forward and the reverse transitions. In addition, we showed the transition states of these transitions are broad and resemble high energy states.

For RNAP, we focus on the conformational changes of RNAP and DNA in promoter melting process. Using BD, we show that DNA conformation changes in promoter melting occur in three steps. We also show that internal dynamics of RNAP is relevant to facilitate the bending of DNA. For Pol II, the structural transitions between two initiation conformational states and between initiation state and elongation state are studied using SPM and BD. We determine the structural units that regulate structural transitions and describe the transition kinetics. The combination of three different methods, SCA, SPM and BD, provide results that are in accord with many experiments. Moreover, our description of the detailed structural transitions in these enzymes lead to new insights and testable predictions in these extraordinarily important enzyme functions.

Link between dynamics and function in single and multi-subunit enzymes

by

Jie Chen

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2010

Advisory Committee:
Professor Devarajan Thirumalai, Chair/Advisor
Professor Sergei Sukharev
Professor John D. Weeks
Professor Arthur La Porta
Professor Dorothy Beckett

Acknowledgments

I want to thank many people in bringing my dissertation research in the past five years to a successful completion, because of whom my graduate experience has been one that I will cherish forever.

First and foremost I'd like to thank my advisor, Professor Devarajan Thirumalai for giving me an invaluable opportunity to work on several challenging and interesting projects over the past five years. He has constantly encouraged and supported me throughout my dissertation research. While giving me large degree of freedom to pursue interesting scientific problems, he always made himself available for help and advice. Dave was very helpful in discussing options for my future career path, including post-doctoral positions and beyond. I learned from him many things including how to select research topics, how to write good research papers, and how to be a scientist. It has been a real pleasure to work with and learn from such an extraordinary individual.

I must thank my colleagues in Thirumalai group who have enriched my graduate life in many ways and deserve a special mention. I would like to thank Dr. Dima Ruxandra (now an assistant professor at University of Cincinnati), whose help and guidance as a mentor in the first year of my research helped me immensely. She taught me the statistical coupling analysis method introduced in Chapter 1. I would like to thank Dr. Changbong Hyeon (now a professor at Korea Institute for Advanced Study), whose unflagging willingness to discuss research ideas and help in computational modeling are valued. Thanks are also due to Dr. Margaret Shun

Cheung (now an asistant professor at University of Houston), who is like a big sister to me, and has offered me not only invaluable research ideas but also life advice, Dr. Subramanian Vaitheeswaran for helping me with cluster computing, Dr. Edward P. O'Brien for building up coarse-grained models for proteins, Dr. Greg Morrison, Dr. Rinna Tehver, and Dr. Sam Cho for proofreading and offering valuable advice on several applications. My interaction with Toan M. Ngo, Zhenxing Liu, Govardhan Reddy, Natalia A. Denesyuk, and Lance(Zhechun) Zhang has been fruitful. I would also like to thank them for giving useful advice for my postdoctoral interview talks.

I would like to acknowledge help and support from our staff members Debbie Jenkins and Caricia Fisher, their help is highly appreciated.

Last but not least, I thank Prof. Sergei Sukharev, Prof. John D. Weeks, Prof. Arthur La Porta, and Prof. Dorothy Beckett for agreeing to serve on my thesis committee.

I owe my deepest thanks to my family - my mother and father who have always stood by me and guided me, and have pulled me through against impossible odds at times. Words cannot express the gratitude I owe them. I would also like to thank my husband, Lei Zhou, this thesis would not be possible without him.

Table of Contents

List of Figures vii

List of Abbreviations ix

1 Introduction 1
1.1 Biopolymer dynamics and relation to the functions 1
1.2 Allosteric wiring diagram (AWD) from statistical coupling analysis (SCA) 5
1.2.1 Statistical coupling energies 5
1.2.2 Perturbation of the MSA 6
1.2.3 Similarity measures 7
1.2.4 Clustering algorithm 10
1.3 Structural perturbation method (SPM) 12
1.4 Brownian dynamics simulation 13
1.4.1 Self-organized polymer (SOP) model 13
1.4.2 Brownian dynamics simulations of conformational changes . . 15
1.4.3 Procedure for inducing C→O transition 17
1.5 Thesis outline and summary 18

2 Dihydropholate Reductase and allosteric signal transduction 21
2.1 Overview 21
2.2 Conformation fluctuations in DHFR 22
2.3 Allostery wiring diagram (AWD) 27
2.3.1 Key residues predicted by SCA are dispersed throughout the structure 27
2.3.2 AWD obtained using SPM overlap with that obtained from SCA 30
2.4 Kinetics analysis of DHFR signaling pathways 32
2.4.1 Anticorrelated Motions of DHFR in the CS and OS states . . 32
2.4.2 Deformation of the Met20 loop in CS→OS transition 35
2.4.3 Sliding of the Met20 loop across $\alpha 2$ limits the CS→OS transition rate 39
2.4.4 Cysteine crosslink inhibits CS→OS transition 43
2.4.5 Deformation of the Met20 loop in the OS→CS transition . . . 44
2.4.6 Transition state emsemble 46
2.5 Residues in the AWD code for ligand binding and dynamics 48
2.6 A small fraction of OS (CS) is present under equilibrium conditions in the CS (OS) state 49
2.7 Concluding Remarks 52
2.8 Method 55

3 Promoter melting triggered by bacterial RNA polymerase 58
3.1 Overall goals: . . . 58
3.2 Transcription cycle of RNAP . . . 59
3.3 Structures of RNAP . . . 61
3.4 Experimental results . . . 64
3.5 Theoretical approaches . . . 66
3.5.1 A network of contacts trigger the promoter melting. . . . 68
3.5.2 $R{\cdot}P_c \rightarrow R{\cdot}P_o$ transition trajectories partition into fast (efficient) and slow (inefficient) tracks: . . . 69
3.5.3 Efficient melting . . . 70
3.5.4 Inefficient melting . . . 71
3.5.5 Three steps in transcription bubble formation . . . 74
3.5.6 Scrunching is a universal mechanism in initiation: . . . 77
3.5.7 Active channel widening . . . 80
3.5.8 Structural fluctuations are enhanced in the slow trajectories: 85
3.5.9 Deletion mutation study shows that removing S results in partial formation of the $\mathrm{R \cdot P_o}$ . . . 87
3.6 Concluding remarks . . . 89
3.7 Methods . . . 90
3.7.1 Rationale of the Hamiltonian switching method . . . 90
3.7.2 Technical details on triggering the conformational changes . . 92
3.7.3 Persistence length of DNA . . . 95
3.7.4 Energy changes of RNAP . . . 97
3.7.5 Free energy profile . . . 98

4 Eukaryote RNA polymerase II and the conformational changes involved in transcription 100
4.1 Overview . . . 100
4.2 Architecture of Pol II . . . 102
4.3 NMA analysis: Pol II holoenzymes and TEC . . . 105
4.3.1 Three intrinsic motions trigger the conformational transitions of Pol II. . . . 106
4.3.2 Function-related structural units that promotes intrinsic motions are identified using the structural perturbation method (SPM). . . . 108
4.4 Brownian dynamics simulation of the conformational transitions . . . 111
4.4.1 The global motion of Pol II is dominated by the motion of the clamp module throughout the transitions. . . . 111
4.4.2 The clamp motion results in narrowing/widening of the DNA channel in F1→F2 (F2→EC) transition. . . . 112
4.4.3 Formation of multiple native contacts between the clamp and shelf modules in the F1→F2 transition triggers the back-forth motion. . . . 114
4.4.4 Ordering of multiple switches in F2→EC transitions in the active center. . . . 117

4.5 Concluding Remarks 119

Bibliography 120

List of Figures

2.1 Catalytic cycle of DHFR. . . . 23
2.2 Sequence conservation of DHFR. . . . 28
2.3 Allostery wiring diagram (AWD) of DHFR obtained using SPM . . . 31
2.4 Correlated and anti-correlated motions of DHFR. . . . 35
2.5 Global RMSD during the CS→OS transition. . . . 36
2.6 Global and local RMSD. . . . 38
2.7 Rupture of contacts during the CS→OS transition. . . . 39
2.8 Sliding motion of Met20 loop. . . . 40
2.9 Sliding motion in DHFR. . . . 42
2.10 RMSD of the reverse transition. . . . 44
2.11 Transition state ensemble. . . . 47
2.12 RMSD for crosslinking mutant during the CS→OS transition . . . 50
2.13 Pre-equilibrition of CS and OS state DHFR . . . 51
2.14 Multiple sequence alignment of DHFR family . . . 56
2.15 Critical size of the MSA . . . 57

3.1 Transcription Cycle of RNAP . . . 60
3.2 RNA polymerase is a crab-claw shape enzyme . . . 61
3.3 Coenzyme: σ factor . . . 62
3.4 DNase footprint experiment . . . 65
3.5 Structural models for the promoter DNA and RNAP. . . . 67
3.6 Contact maps of RNAP complex with DNA . . . 69
3.7 RMSD shows the efficient and inefficient melting processes . . . 72
3.8 Efficient and inefficient melting processes . . . 73
3.9 Inefficient melting . . . 74
3.10 Three-step melting process . . . 78
3.11 Scrunching mechanism . . . 79
3.12 Rope-swing mechanism . . . 82
3.13 Distance changes upon channel opening . . . 84
3.14 Molecular details of channel opening mechanism for a slow trajectory. 86
3.15 Removal of S leads to partial formation of $R \cdot P_o$ . . . 87
3.16 Hamiltonian switching method . . . 91
3.17 Effects of varing switching condition . . . 94
3.18 Distribution of DNA end to end distances . . . 96
3.19 Potential energy of RNAP-DNA complex . . . 97
3.20 Free energy profile . . . 99

4.1 Crystal structures of Pol II . . . 103
4.2 Four modules in Pol II . . . 104
4.3 NMA results . . . 107
4.4 SPM analysis . . . 109
4.5 RMSD as a function of time . . . 113
4.6 Global motion in Pol II . . . 115

4.7 The evolution of distances of multiple contacts 116
4.8 The evolution of distances of multiple salt-bridges 118

List of Abbreviations

AWD	Allosteric Wiring Diagram
BD	Brownian Dynamics
C	Closed
CL	Crosslink
CS	Closed State
CSE	Chemical Sequence Entropy
CTWC	Coupled Two-Way Clustering
DHF	7,8-Dihydropholate
DHFR	Dihydrofolate Reductase
EC	Pol II-DNA Elongation Complex
F1	Form 1 of 10-Subunit Pol II Core Enzyme
F2	Form 2 of 10-Subunit Pol II Core Enzyme
FENE	Finite Extensible Nonlinear Elastic
FRET	Frster Resonance Energy Transfer
H	Hydrophobic
H2	Helix 2 of DHFR
MFPT	Mean First Passage Time
MSA	Multiple Sequence Analysis
MT	Magnetic Tweezers
MTX	Methotrexate
NADP	Nicotinamide Adenine Dinucleotide Phosphate
NMA	Normal Mode Analysis
NMR	Nuclear Magnetic Resonance
NT	Non-template
O	Open
OS	Open State
P	Polor
PDB	Protein Data Bank
Pol II	RNA Polymerase II
PSI	Position Specific Iterative
R	Rope Domain
RMSD	Root Mean Square Deviation
RNAP	Bacterial RNA Polymerase
Rpb1	Subunit 1 of Pol II
Rpb5	Subunit 5 of Pol II
$R{\cdot}P_c$	RNAP-Promoter DNA Closed Complex
$R{\cdot}P_{itc}$	RNAP-Promoter DNA Initiation Transcription Complex
$R{\cdot}P_o$	RNAP-Promoter DNA Open Complex
RS	Rope-Swing

S	Swing Domain
SCA	Statistical Coupling Analysis
SOP	Self-Organized Polymer
SPC	Superparamagnetic Clustering
SPM	Structural Perturbation Method
SW	Swendsen-Wang
Taq	*Thermus Aquatics*
TEC	Transcription Elongation Complex
THF	Tetrahydrofolate
TS	Transition State
TSE	Transition State Ensemble

Chapter 1

Introduction

1.1 Biopolymer dynamics and relation to the functions

Polymers in solutions incessantly change both their shape and position randomly by thermal agitation[1]. For biopolymers, proteins and nucleic acids, these brownian motions dominate biological functions such as substrate binding, product release, allosteric regulation, and motor functions. At a microscopic level, atoms in biopolymers and their environment evolve towards equilibrium due to thermal motions. However, the equilibrium can be changed by the external conditions - pH, temperature, concentration - and shifts to some other states. As a result, biopolymers undergo large conformational changes to evolve towards their new equilibrium states. These conclusions also come from the accepted notion that the energy landscape of enzymes even in the folded state is rugged [2], and hence thermal energy might be sufficient to access several conformational substates during a typical reaction cycle [3]. The thermal-fluctuation driven conformational changes have been experimentally observed, and a direct link between μs to ms protein domain motions and enzymatic functions has been established [4, 5, 6, 7]. However, the connection between the functional related, collective motions with local fluctuations, which are on ps to ns time scale, is largely unknown [4, 8, 5]. A great challenge we face is to understand the impact of molecular motions within the protein on the enzyme's

catalytic properties[9, 10].

A naive scenario of dynamics for enzymes is based on two global time scales: τ_E, which is the time scales for protein conformational changes, and τ_F, which is the time scale for the "chemical reaction" in the enzyme. When $\frac{\tau_E}{\tau_F} >> 1$, the conformational changes of an enzyme are much slower than the chemical reaction catalyzed by the enzyme, therefore, the conformational changes are not directly coupled to the enzyme catalytic function. Complex reaction coordinates have to be selected to incorporate motions of different time scales, and generate the appropriate free energy profile. On the other hand, if $\frac{\tau_E}{\tau_F} \approx 1$ the conformational changes occur in the same time scale as the chemical reaction. The conformational fluctuations are directly coupled to the catalytic process, and free energy profile can be easily explored with simple reaction coordinates such as the enzyme structure.

The link between conformational dynamics and enzyme catalysis has been extensively studied for several enzymes where ligand binding regulates allostery, which in turn controls the nature of catalysis. Regulation by allosteric enzymes involve binding of effector molecules and changes of protein conformation in sites distant to effector binding site [11, 12]. Dihydrofolate reductase (DHFR), is an example of a well-studied allosteric enzyme with $\frac{\tau_E}{\tau_F} >> 1$, catalyzes hydride transfer between nicotinamide adenine dinucleotide phosphate (NADP) and 7,8-dihydrofolate (DHF). During the reaction cycle [13], DHFR undergoes conformational changes between closed (CS) and occluded states (OS). In recent years, experiments and simulations ([14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24]) point to the fact that the reaction coordinate is complex, involving a network of collective motions of active site and parts

of the enzyme distant from the active site. Illustration of the dynamics of the entire enzyme is needed to obtain new insights into the understanding of the biological function of DHFR [6, 16, 17].

We perform analysis of the complex link between structure, movement, and catalysis of an enzyme using a number of new methods based on sequence, structure, and dynamical simulation. Statistical coupling analysis (SCA) is based on the evolutionary information stored in protein sequences, and can be used to extract the evolutionarily conserved and co-evolved amino acid positions in a protein family through sequence statistics and coupling analysis [25, 26, 27]. It is based purely on evolutionary information, which is ideal for identifying a network of residues that facilitate allosteric transitions between various distinct states (CS and OS in DHFR).

A structure based method, Normal mode analysis (NMA), is based on the assumption that, in equilibrium, energy surface can be characterized by harmonic approximation [28, 29, 30, 31, 32]. Despite the approximations, NMA has been found to be useful to determine the function-relavent harmonic modes [29, 30, 31, 32]. We introduced a new structural perturbation method (SPM) based on NMA, to extract residues that caused large changes on the function-related motions in a protein [30, 31, 32, 33].

Besides the static pictures of the allosteric wiring diagram that emerges from the SPM, we are most interested in the transition dynamics of a protein, which are difficult to obtain using experiments alone. Using molecular dynamics simulations, information of the energy landscape explored during the allosteric transitions can be

extracted. The questions of interest are: 1) What are the pathways connecting these conformational states? 2) What are the structures of the transition state ensemble (TSE)? These questions can be answered using Brownian dynamics simulation (BD), and a coarse-grained self-organized polymer (SOP) model. The coarse-grained models allows us to sample the energy landscape exhaustively, thus yielding reliable estimates of thermodynamics as well as kinetics [34, 35]. For the allosteric transition of DHFR CS→OS and the reverse transition, we can easily perform $\mu s \sim ms$ BD to study the transition dynamics.

BD with SOP model is especially powerful for studying large biological systems such as motor proteins and molecular machines. The size of these biopolymers invokes long time scales ($\mu s \sim ms$) of the conformational dynamics and makes it impossible to perform traditional all-atom molecular dynamics simulations. Besides, we are not interested in the detailed oscillations of atoms around the equilibrium on the ps to ns time scales, but in the large scale motions of the biopolymers. Therefore, coarse-graining large biopolymers using SOP is ideal for our purpose. We have establish the link between dynamics and function for a molecule machine, RNA polymerase, which is responsible for RNA synthesis using DNA as a template. Using BD and SOP for protein and DNA, we can study the complex nature of the promoter melting process, which involves RNAP induced melting of the promoter DNA by ± 12 base pairs, and bending of it into the active site channel of RNAP. Although a number of methods, biochemical experiments [36, 37, 38, 39], crystallography [40, 41], and single molecule approaches [42, 43, 44], have been used to try to understand this process, our dynamic simulations unveil the promoter melting process in the

transcription initiation for the first time.

In the following section, we start with introducing the basics of SCA and SPM methods. Then we give the theoretical framework of performing BD simulation of conformational transitions. In section 1.5, we provide a summary of the thesis.

1.2 Allosteric wiring diagram (AWD) from statistical coupling analysis (SCA)

SCA is a statistical method for extracting the evolutionarily conserved and co-evolved amino acid positions in a protein family based on multiple sequence alignment. There are three key ingredients in this method: 1) defining statistical meaningful coupling energies 2) defining reliable similarity measures 3) choosing an efficient clustering method. In this section, we will introduce SCA based on these key ingredients.

1.2.1 Statistical coupling energies

Based on the multiple sequence alignment (MSA) of a protein family, a statistical free energy for each amino acid position, ΔG_i, is defined as the overall deviation of amino acid frequencies at the position i from their mean values [45],

$$\frac{\Delta G_i}{kT^*} = \sqrt{\frac{1}{C_i}\sum_{x=1}^{20}[p_i^x ln(\frac{p_i^x}{p_x})]^2}. \tag{1.1}$$

Here, kT^* is an arbitrary energy unit and C_i is the number of types of amino acid that appears at position i, p_x is the mean frequency of amino acid type x in the MSA. $p_i^x = \frac{n_i^x}{N_i}$, where n_i^x is the number of amino acid type x appears at position

i and N_i is the total number of valid amino acid at position i excluding gaps. This definition of free energy is very similar to the definition of sequence entropy $S_i = -\sum_{x=1}^{20} p_i^x ln(p_i^x)$. For a given MSA, we used a cutoff of $\Delta G_i > 10$ to define conserved residues, which is equivalent to $S_i < 1$.

Another important definition is correlation of two positions i, j in MSA, which is given by the change of amino acid frequencies at position i as the consequence of perturbation of the MSA at position j,

$$\frac{\Delta\Delta G_{ij}}{kT^*} = \sqrt{\frac{1}{C_i}\sum_{x=1}^{20}[p_i^{x\prime} ln(\frac{p_i^{x\prime}}{p_x}) - p_i^x ln(\frac{p_i^x}{p_x}]^2} \tag{1.2}$$

where $p_i^{x\prime} = \frac{n_i^{x\prime}}{N_i}$, $n_i^{x\prime}$is the number of amino acid x appear at position i in the perturbed MSA and $N_i' = \sum_{x=1}^{20} n_i^{x\prime}$.

1.2.2 Perturbation of the MSA

Perturbations of the MSA are done by building subsets of MSA. If position j is perturbed, the corresponding subset of MSA is gotten by retaining sequences which have the specific type of amino acid, x, at the position j, where x is the amino acid which appears most frequently at position j. It is obvious that the amino acid frequencies at position i would deviate greatly from the original MSA if position i and j are correlated under evolutionary pressure. On the other hand, $\Delta\Delta G_{ij}$ is 0 if 1) j is a perfectly conserved position ($p_i^x = p_i^{x\prime}$); 2) j is not correlated with i and all amino acid at i are found at their mean frequencies in the MSA ($\frac{p_i^x}{p_x} = \frac{p_i^{x\prime}}{p_x} = 1$).

It is worth noting that the size of subset MSA should be large enough to represent of statistical properties of the original MSA. Following Dima and Thirumalai

[45], central limit theorem is used to determine the size of sub-alignment that contain $P = N \cdot f$ sequences. The number of alignments for a fixed f is $k = \frac{N!}{P!(N-P)!}$. To obtain statistically meaningful results, the general properties of the subalignments must be similar to the original MSA. In analogy with statistical mechanics, we suggest that the smallest value of f be chosen so that the law of large numbers is obeyed. In particular, we choose f so that the following criteria are satisfied:

$$\langle \overline{\Delta G} \rangle_f = \frac{\sum_{l=1}^{k} \overline{\Delta G_l}}{k} \approx \overline{\Delta G}_{MSA} \tag{1.3}$$

$$\sigma_f^2 = \langle \overline{\Delta G_l}^2 \rangle_f - \langle \overline{\Delta G} \rangle_f^2 \tag{1.4}$$

where $\overline{\Delta G_l} = \frac{1}{L_{MSA}} \sum_{i=1}^{L_{MSA}} \Delta G_i^l$, $\overline{\Delta G}_{MSA} = \frac{1}{L_{MSA}} \sum_{i=1}^{L_{MSA}} \Delta G_i$, σ_f is the width of the distribution of $\overline{\Delta G_l}$. Failure to satisfy these criteria can give spurious results in the applications of SCA.

1.2.3 Similarity measures

The matrix $\mathbf{G}$, whose elements are the $\Delta\Delta G_{ij}$ values for a protein family, represents the response of positions i in the MSA to all allowed perturbations at site j provided the perturbations satisfy the acceptance criteria stated above. The rows of the matrix correspond to positions in the MSA and the columns to perturbations. In order to reliably determine the network(s) of positions that change in a correlated manner starting from this matrix, we used the coupled two-way clustering (CTWC) that was developed to analyze DNA-microarray data [46]. The basic idea is to carry out successive elementary rounds of Superparamagnetic clustering (SPC) [47]. At each step, the submatrix that contains positions and perturbations that cluster

together in the previous iteration with large signals is extracted. An important ingredient in the SPC technique is the choice of a similarity measure between a pair of entries that are to be clustered. In the context of clustering of positions in an MSA, there are at least two natural choices for similarity measures [45], (1) the Euclidean distance and (2) the Pearson correlation coefficient. In what follows, we give the rationale and the details for using these measures. The collection of the $\Delta\Delta G_{ij}$ values (with j varying from 1 to L_{MSA} with L_{MSA} being the total number of positions, including gaps, of the alignment) for a given position i in the MSA can be thought of as a vector with L_{MSA} components, $\overrightarrow{v_i} = \{\Delta\Delta G_{i1}, \cdots, \Delta\Delta G_{iL_{MSA}}\}$. Therefore, the degree of similarity between two positions i and k can be represented by the Euclidean distance between the two corresponding vectors, i.e.,

$$D_{ik} = \sqrt{\sum_{j=1}^{L_{MSA}} (\Delta\Delta G_{ij} - \Delta\Delta G_{kj})^2} \tag{1.5}$$

For each MSA there is a spread in the magnitudes of the $\Delta\Delta G_{ij}$ values (e.g., from ~ 0.01 to ~ 10). Thus, for a pair of small matrix elements, D_{ik} will be small even if the two vectors are not similar. On the other hand, for two related positions i and k with large $\Delta\Delta G_{ij}$ values, a difference in any of their components could lead to a large D_{ik} value that would not reflect their true similarity. Positions with small $\Delta\Delta G_{ij}$ values are of little interest because they show basically no response to changes in other positions in the MSA.

To correct for the potentially spurious results indicated above, we use the following protocol: (1) We eliminate entries (positions) that show virtually no response to the overwhelming majority of perturbations. (2) We scale the $\Delta\Delta G_{ij}$ values so

that only a few categories of the matrix elements are included in the analysis. To a large extent, the results do not depend on the precise boundaries used in the classification of $\Delta\Delta G_{ij}$. (3) The D_{ik} values are suitably normalized. If all or all but one of the corresponding $\Delta\Delta G_{ij}$ values are < 1.0, then the row corresponding to position i is deleted from the input data matrix. The scaling of the $\Delta\Delta G_{ij}$ values is achieved by assigning them to two or three characteristic entries. For example, all the small $\Delta\Delta G_{ij}$ values (i.e., $\Delta\Delta G_{ij}$) are kept unchanged, while the intermediate $\Delta\Delta G_{ij}$ values (i.e., $1.0 \leq \Delta\Delta G_{ij} < 2.0$) are assigned a value α_1 and the remaining (large) $\Delta\Delta G_{ij}$ values are assigned a value α_2 such that $\alpha_1 \sim 10$ and $\alpha_1 < \alpha_2$. We normalize D_{ik} using

$$SE_{ik} = \frac{D_{ij}}{0.5 \times (|\overrightarrow{v_i}| + |\overrightarrow{v_k}|)} \tag{1.6}$$

where $|\overrightarrow{v_i}|$ is the norm of the vector $\overrightarrow{v_i}$. SE_{ik} is small for pairs of vectors that have components of similar values, and it is independent of the actual magnitude of the individual components. In addition, because positions that show reduced or no response to the majority of the perturbations are eliminated, a small SE_{ik} value indicates that the two positions show large responses to the same set of perturbations. A second similarity measure that can be used is the Pearson correlation coefficient

$$P_{ik} = \frac{\sum_{j=1}^{L_{MSA}} (\Delta\Delta G_{ij} - \langle\Delta\Delta G_i\rangle)(\Delta\Delta G_{kj} - \langle\Delta\Delta G_k\rangle)}{\sigma_i \sigma_k} \tag{1.7}$$

where $\langle\Delta\Delta G_i\rangle = \frac{\sum_{j=1}^{L_{MSA}} \Delta\Delta G_{ij}}{L_{MSA}}$ is the average $\Delta\Delta G_{ij}$ value for position i and $\sigma_i^2 = \sum_{j=1}^{L_{MSA}} (\Delta\Delta G_{ij} - \langle\Delta\Delta G_i\rangle)^2$ is the variance. Just as with the Euclidean distance similarity measure, P_{ik} is small for two positions with little or no responses to the majority of perturbations. Prior to calculating P_{ik}, we eliminate all such positions.

The procedure used is the same as described above. For two perfectly correlated (anticorrelated) positions, $P_{ik} = 1(-1)$, while for uncorrelated positions, $P_{ik} = 0$. Because the Euclidean distance measure (SE_{ik}) has small values for two correlated positions and we want to be able to use the two similarity measures interchangeably, we replaced P_{ik} with

$$SP_{ik} = 1 = |P_{ik}| \tag{1.8}$$

where $|P_{ik}|$ is the absolute value of P_{ik}. Therefore, both SE_{ik} and SP_{ik} are zero when the two positions are perfectly correlated.

The Euclidean similarity measure SE_{ik} is best suited when the individual $\Delta\Delta G_{ij}$ values are not broadly distributed, that is, when the largest $\Delta\Delta G_{ij}$ value is ~ 3.0. In such a case, the responses of each position in the alignment to the various perturbations are similar in magnitude. Therefore, the use of the Pearson correlation coefficient SP_{ik} would lead to the majority of positions being clustered. On the other hand, using the SE_{ik} and the associated rescaling of entries allows us to distinguish between positions, and therefore only a handful of positions turn out to be clustered after the application of the CTWC procedure. It follows then that the Pearson similarity measure is best suited for MSA for which there is a broad distribution in the magnitude of the responses of positions to perturbations.

1.2.4 Clustering algorithm

The Swendsen-Wang cluster Monte-Carlo algorithm (SW) is used to cluster the data points (positions or perturbations in $\Delta\Delta G$ matrix) so that the data points in

the same cluster are similar to each other than to points in different clusters. Based on an inhomogeneous Potts model, a Potts spin is assigned to each data point and short range interactions between neighboring points are introduced. Spin-spin correlations (measured by Monte-Carlo computations) serve to partition the data points into clusters. The clustering procedure consists two main steps: First we identify the range of temperatures where the clusters appear. Secondly, at temperatures that are within this range, the clusters are identified [48]. In the following, the procedure is summarized: 1) Assign to each data point i a q-state Potts spin s_i (we choose $q = 20$). 2) Find the K-order nearest neighbors (n.n.) for each data point based on similarity measure given above (K varies between 10 and 20). 3) Measure the nearest neighbor interactions $J_{ij} = J_{ji} = \frac{1}{\hat{K}} exp(-\frac{SE_{ik}}{2a^2})$, where $\hat{K}$ is the average number of neighbors per date point, and a is the average distance between all n.n. . 4) Using the SW algorithm and the Hamiltonian $H[\{s\}] = -\sum_{\langle i,j \rangle} J_{ij} \delta_{s_i,s_j}$ to calculate susceptibility $\chi = \frac{N}{T}(\langle m^2 \rangle - \langle m \rangle^2)$. Here $m = \frac{(N_{max}/N)q-1}{q-1}$, $N_{max} = max\{N_1, N_2, \cdots, N_q\}$ and N_μ is the number of spins with the value μ. 5) Identify the range of temperatures corresponding to the superparamagnetic phase, between T_{fs}, the temperature of maximal χ, an the (higher) temperature T_{ps} where χ diminishes abruptly. Cluster assignment is performed at $T_c = \frac{T_{fs}+T_{ps}}{2}$. 6) Measure at $T = T_c$ the spin-spin correlation function $\langle \delta_{s_i,s_j} \rangle$ for all pairs of neighboring points. 7) Clusters are identified as the data points that satisfy $\langle \delta_{s_i,s_j} \rangle > \theta$, here $\theta = 0.5$, and all mutual friends are assigned to the same cluster.

1.3 Structural perturbation method (SPM)

SPM is developed to assess the dynamics importance of each residue of a protein based on normal mode analysis (NMA) [30, 31]. The basic idea of NMA is to build an elastic network in which harmonic potentials are used for the pairwise interactions between all C_α atoms within a cutoff distance ($R_c = 10\mathring{A}$). The Hamiltonian of the network is:

$$E_{network} = \frac{1}{2} \sum_{d_{ij}^0 < R_c} C(d_{ij} - d_{ij}^0)^2 \tag{1.9}$$

where d_{ij} is the distance between residues i and j, and d_{ij}^0 is the same distance in the crystal structures. We choose a single force constant $C = 10kcal/(mol\mathring{A}^2)$ for the network.

By performing standard NMA with Eq. (1.9), we get a set of eigenvalues and eigenvectors, among which the low normal modes usually correlate to function-related motions [29]. For each of the mode m, we overlap the corresponding eigenvector v_{im}, with the conformational changes, $\Delta r_i = r_i^C - r_i^O$, here, r_i^C and r_i^O are positions of residue i in the closed and open state structures. We calculate overlap using equation [49, 29]:

$$I_m = \frac{|\sum_{i=1}^{3N} v_{im} \Delta r_i|}{|\sum_{i=1}^{3N} v_{im}^2 \sum_{i=1}^{3N} \Delta r_i^2|^{1/2}} \tag{1.10}$$

The transition-related normal mode is identified as the ones that overlap best with the conformational changes. For the selected mode, the dynamic importance of residue i can be assessed by the response of a local perturbation at i. In the context of normal mode analysis, perturbation is realized by small changes in the force constant of those springs that connected to residue i. The response is in terms of

a normalized score $\delta\omega_i^m$, given by $\delta\omega_i^m = V_m^T \Delta H V_m$, where V_m is the eigenvectors of mode m and ΔH is the perturbed Hessian matrix. $\delta\omega_i^m$ is proportional to the changes of elastic energy of residue i, δE_i, which can be calculated using,

$$\delta E_i = \frac{1}{2} \sum_{d_{ij}^0 < R_c} \Delta C (d_{ij} - d_{ij}^0)^2 \tag{1.11}$$

where j is the index of the residue that contacts with i and $\Delta C = -C/2$.

1.4 Brownian dynamics simulation

1.4.1 Self-organized polymer (SOP) model

The long-time scales invoked in the conformational changes in the catalytic cycle make it necessary to use a coarse-grained self-organized polymer (SOP) model [35, 50] for efficient sampling energy landscape. In the SOP model, the structure of a protein is represented using only the C_α coordinates, $r_i^P (i = 1, 2, ...N^P)$ with N^P being the number of amino acid), while the structure of a nucleotide is represented by the centers of mass, $r_i^N (i = 1, 2, ...N^N)$ with N^N being the number of nucleotides. The state-dependent energy function in the SOP representation for a protein-nucleic complex (energy functions for protein or nucleic acid alone can be easily derived) is

$$\begin{aligned} H(\{r_i\}|X) &= V_{FENE} + V_{NB}^{ATT} + V_{NB}^{REP} + V_{NB}^{ELEC} \\ &= -\sum_{Y=P,N} \sum_{i=1}^{N^Y-1} \frac{k}{2} R_0^2 \log(1 - \frac{(r_{i,i+1}^Y - r_{i,i+1}^{Y,0}(X))^2}{R_0^2}) \\ &+ \sum_{Y=P,N} \sum_{i=1}^{N^Y-3} \sum_{j=i+3}^{N^Y} \epsilon_h^Y [(\frac{r_{ij}^{Y,0}(X)}{r_{ij}^Y})^{12} - 2(\frac{r_{ij}(X)^{Y,0}}{r_{ij}^Y})^6] \Delta_{ij} \\ &+ \sum_{i=1}^{N^P} \sum_{j=1}^{N^N} \epsilon_h^{P,N} [(\frac{R_{ij}^0(X)}{R_{ij}})^{12} - 2(\frac{Rij^0(X)}{R_{ij}})^6] \Delta_{ij} \end{aligned} \tag{1.12}$$

$$+ \sum_{Y=P,N} \Big(\sum_{i=1}^{N^Y-2} \epsilon_l^Y \Big(\frac{r_{i,i+2}^{Y,0}}{r_{i,i+2}^Y}\Big)^6 + \sum_{i<j} \epsilon_l^Y \Big(\frac{\sigma^Y}{r_{ij}^Y}\Big)^6 (1-\Delta_{ij})\Big)$$
$$+ \sum_{i=1}^{N^P} \sum_{j=1}^{N^N} \epsilon_l^{P,N} \Big(\frac{\sigma^{P,N}}{R_{ij}}\Big)^6 (1-\Delta_{ij})$$

The label X denotes conformational states of a protein or nucleic acid, and Y refers to protein (P) or nucleic acid (N). In Eq. (1.12), $r_{i,i+1}^Y$ is the distance between two adjacent C_α-atoms (centers of mass of nucleotides) and $r_{i,j}^Y$ gives the distance between the i^{th} and j^{th} C_α-atoms (centers of mass of nucleotides) and $R_{i,j}$ is the distance between the i^{th} C_α-atom in the enzyme and center of mass of the j^{th} nucleotide in nucleic acid. The superscript "o" denotes their values in state X.

The first term in Eq. (1.12), the finite extensible non-linear elastic (FENE) potential, accounts for chain connectivity. The stability of the state X is described by the non-bonded interactions (the second term in Eq.(1.12)) that assigns attractive interaction between two residues or nucleotides that are in contact in X. Non-bonded interactions between residues or nucleotides that are not in contact in X are taken to be purely repulsive (the third term in Eq. (1.12)). The value of $\Delta_{i,j}$ is 1 if i and j are in contact in state X, and is zero otherwise. A native contact implies that the distance between the i^{th} and j^{th} centers is less than a cut-off distance R_C.

The spring constant, k, in the FENE potential (the first term in Eq.(1.12)) for stretching a covalent bond is 20 kcal/(mol$\mathring{A}^2$), and the value of R_0, which gives the allowed extension of the covalent bond, is 2 $\mathring{A}$. The values of the parameters for protein and nucleic acid and for the protein-nucleic acid interactions, which are

obtained by using the standard combining rules [51],

$$\epsilon = \sqrt{\epsilon_1 \epsilon_2} \tag{1.13}$$

$$\sigma = \frac{\sigma_1 + \sigma_2}{2}$$

are given in table 1. It is worth noting that there are only a few parameters in the SOP, which allows us to fully explore the fundamental physical processes of biopolymers.

Table 1.1: Parameters for SOP model for protein and nucleic acid

		Protein	Nucleic acid	Protein-Nucleic acid
R_0	$(\mathring{A})$	2	2	2
k	$(kcal/(mol \cdot \mathring{A}^2)$	20	20	20
R_C	$(\mathring{A})$	8	14	11
ϵ_h	$(kcal/mol)$	2	0.7	1.2
ϵ_l	$(kcal/mol)$	1	1	1
σ	$(\mathring{A})$	3.8	7	5.4
ζ	(τ_L^{-1})	50	50	50
τ_L	(ps)	3	3	3

1.4.2 Brownian dynamics simulations of conformational changes

For proteins that change conformation between closed (C) and open (O) states, the kinetics of the transitions C→O are probed using a method that has been successfully used to study the allosteric transitions in GroEL[52]. The basic assumption of the method is that the local strain accumulated during ligand or cofactor binding propagates faster than the global conformational transitions.

Using SOP model, the transition C→O (or the reverse transition) can be simulated by assuming that the dynamics of the system interested in can be adequately

described by the Langevin equation in the overdamped limit. We start from the C state by performing equilibrium simulations using Eq.(1.12) with X being the C state. Then, the transition to the O state is induced by using the forces computed from the energy function $H(\{\vec{r}_i\}|O)$. The Langevin equations of motion for C→O are,

$$-\zeta\frac{\partial\vec{r}_i}{\partial t} = -\frac{\partial H\{\vec{r}_i|O\}}{\partial\vec{r}_i} + \vec{\Gamma}_i(t), \tag{1.14}$$

where ζ is the friction coefficient, $\vec{r}_i$ is position of the ith residue (nucleotide) at time t, and $\Gamma_i(t)$ is the random force. The initial (t=0) position of ith residue (nucleotide), r_i, is taken from the Boltzmann distribution at temperature $T = 300K$ corresponding to C,

$$P(\vec{r}_i(0)) \sim e^{-\beta H(\vec{r}_i|C)} \tag{1.15}$$

where $\beta = 1/k_BT$ and k_B is the Boltzmann constant. The random force, $\Gamma(t)$, satisfies

$$< \Gamma(t) >= 0 \tag{1.16}$$

and

$$< \Gamma(t)\Gamma(t') >= 2k_BT\zeta\delta(t-t') \tag{1.17}$$

where the averages are over the trajectories. At long times, the ensemble of conformations obeys the Boltzmann distribution corresponding to the O state, so that $P(r_i(t)) \sim e^{-\beta H(r_i|O)}$. Thus, as long as the standard potential conditions are satified [53], the conformational changes can be realized and the microscopic events that drive the transition can be described from the ensemble of trajectories that connect the C and O states.

1.4.3 Procedure for inducing C→O transition

The procedure used to induce the transitions is the following. During the very early stages of the transition between the two states, we define $\vec{r}_{ij}^{C\to O}$ to be the combination of $\vec{r}_{ij}^{C}$ and $\vec{r}_{ij}^{O}$,

$$\vec{r}_{ij}^{C\to O} = \frac{(K-k)\cdot\vec{r}_{ij}^{C} + k\cdot\vec{r}_{ij}^{O}}{K}, \tag{1.18}$$

where $K = 100$, and $k = 0$ at the beginning of the switching, which is then increased by one every 1000 time steps up to 100. Thus, the switching of the energy functions from $H(\{\vec{r}_i\}|C)$ to $H(\{\vec{r}_i\}|O)$ occur over $100,000$ time steps. Note that one can vary K and k values to simulate different strain propagation time (see discussions in [52, 54] and section 3.7.2).

Computation of time scales We estimate the time scale involved in the transition using the following method. Using Eq.1.14, when the inertial term dominates, the natural estimate of time is $\tau_L^2 = ma^2/\epsilon_h$. The energy ϵ_h is of the order of $1kcal/mol$, typical values for $m = 3\times10^{-22}g$ for amino acid, and $a = 1\mathring{A}$. We roughly evaluate τ_L to be $0.6ps$. The value of high friction coefficient used in our simulation $\zeta_H = 50m/\tau_L$, which is equal to the ζ_{water} of a sphere with radius $5\mathring{A}$ in water. The natural measure of time for over-damped condition at simulation temperature T_s is

$$\tau_H \approx \frac{\zeta_H a^2}{k_B T_s} = \zeta_H \frac{\tau_L}{m}\times\frac{\epsilon_h}{k_B T_s}\tau_L \tag{1.19}$$

which gives $\tau_H = 50ps$. A single simulation time step is converted to real time using $t = 0.16\tau_H = 8ps$. Therefore, 10^6 time steps correspond to 8 μs. During the

stage when we use Eq. (6) a smaller time step, $t = 0.016\tau_H = 0.8ps$, is chosen to keep the system stable. The mapping of the simulation time to real time should be considered approximate. Previous studies have shown [55, 52, 54] that the estimated time scales are accurate to within an order of magnitude.

1.5 Thesis outline and summary

The theoretical and computational methods described above have been applied to probe enzyme functions and the initial events in the transcription of genetic information by RNA polymerase. In Chap. 2, we present the results of a study of a *Escherichia coli* dihydrofolate reductase (DHFR) which catalyzes the reduction of dihydrofolate (DHF) to tetrahydrofolate (THF). Using sequence-based SCA and structural based SPM methods, we identify a network of residues that represents the allostery wiring diagram of DHFR. Many of the residues in the allostery wiring diagram, which are dispersed throughout the DHFR structure, are not conserved, but shown to be linked to equilibration conformational fluctuations of DHFR. To further probe the nature of the events that occur during the catalytic cycle, we use SOP model to monitor the kinetics of the conformational transitions occur during the allosteric transitions of DHFR. We find that sliding motion of an important Met20 loop is involved in transmitting allosteric signals. Restraining Met20 loop motion impedes the allosteric transitions of DHFR.

In Chap. 3, we present the results for a study of the promoter DNA melting triggered by bacterial RNA polymerase (RNAP). Using coarse-grained SOP model

of *Thermus aquatics* RNAP and DNA complex structures, we performed Brownian dynamics simulations of the promoter melting process. We find that efficient DNA melting trajectories feature local melting of the promoter DNA around the -10 element, which is followed by sequential unzipping of DNA till the +2 site. We find promoter DNA melting occurs in three steps. In step I, dsDNA melts locally In step II, DNA scrunches into RNA polymerase and the downstream base pairs sequentially open to form the transcription bubble, which results in build up of strain. Subsequently, the downstream dsDNA bends and the strain is relieved Entry of the dsDNA into the RNAP requires widening of the active channel, which involves transient movements of a subdomain of the β subunit of RNAP caused by steric repulsion between nucleotides in the template strand and residues in the β subunit of RNAP.

In Chap. 4, we present the results of a study of an eukaryotic RNA polymerase from yeast (Pol II). Two methods, normal mode analysis (NMA) and Brownian dynamics simulation using a Self-organized polymer model (SOP), are used to study the conformational transitions of a 10-subunit yeast RNA polymerase II (Pol II) between two forms of holoenzyme (form 1 and form 2), and between holoenzyme (form 2) and a Pol II elongation complex (EC). NMA analysis for the structures of Pol II holoenzyme and elongation complex reveals open-close, back-forth and outside-in motions of the mobile clamp module. Brownian dynamics simulation provides dynamic pathways of the conformational transitions from form 1 (F1) to form 2 (F2), and form 2 (F2) to elongation complex (EC). In addition to the relative motions of the four mobile modules, we discovered that breakage and/or formation

of mutiple native contacts between Rpb1 and Rpb5 trigger the transtion from F1 to F2, and several salt-bridges are found to be related to the conformational changes of previously identified "switches" in the transition F2 to EC.

Chapter 2

Dihydropholate Reductase and allosteric signal transduction

2.1 Overview

Conformational fluctuations among a network of residues that are dispersed through out the structure of proteins are thought to play a vital role in enzyme catalysis. The importance of such motions has been extensively studied in *E. Coli.* dihydrofolate reductase which catalyzes the reduction of dihydrofolate to tetrahydrofolate. During the catalytic cycle DHFR undergoes conformational transitions between the closed (CS) and occluded (OS) states which, respectively, describe whether the active site is closed or occluded by the Met20 loop (Fig. 2.1). The CS→OS and the reverse transition may be viewed as allosteric transitions. Using a sequence-based approach we first identify a network of residues that represents the allostery wiring diagram (AWD). Many of the residues in the AWD, that are dispersed throughout the adenosine binding domain as well as the loop domain (Fig. 2.1), are not conserved. Several of the residues in the network have been previously shown by NMR experiments [56, 57, 58], mutational studies [59, 18, 60], and molecular dynamics simulations [21, 14, 15] to be linked to catalysis or in the protein motion. To further probe the nature of events that occur during conformational fluctuations we use a self-organized polymer model [50] to monitor the kinetics of the CS→OS and the reverse transitions. During the CS→OS transition, coordinated

changes in a number of residues in the loop domain enable the Met20 loop to slide along the α-helix in the adenosine binding domain. Sliding is triggered by pulling of the Met20 loop by the βG-βH loop and pushing action of the βG-βH loop. The residues that facilitate the Met20 loop motion are part of the network of residues that transmit allosteric signals during the CS→OS transition. Replacement of M16 and G121, whose C_α atoms are about $4.3\mathring{A}$ in the CS, by a disulfide crosslink impedes that CS→OS transitions. The order of events in the OS→CS transition is not the reverse of the forward transition. The contact Glu18-Ser49 in the OS state persists until the sliding of the Met20 loop is nearly completed. The ensemble of structures in the transition state (TS) in both the allosteric transitions are heterogeneous. The most probable TS structure resembles the OS (CS) in the CS→OS (OS→CS) transition which is in accord with the Hammond postulate. Structures resembling the OS (CS) are present as minor ($\sim (1-3)\%$) component in equilibrated CS (OS) structures.

2.2 Conformation fluctuations in DHFR

Conformational fluctuations of proteins have been argued to play a central role in enzyme catalysis [13, 3, 61, 6]. Such a concept is appealing because the energy landscape of enzymes even in the folded state is rugged [2], and hence thermal energy might be sufficient to access several conformational substates during a typical reaction cycle [3]. In recent years, results from a number of studies support the idea that dynamic motions in a network of residues that promote catalytically-relevant

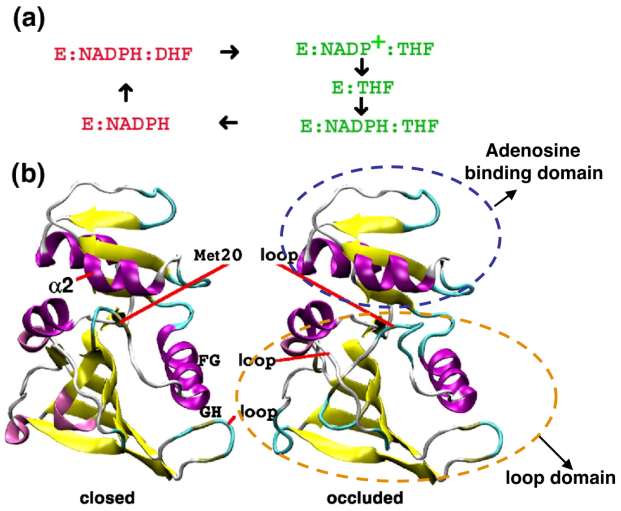


Figure 2.1: Catalytic cycle and structures of closed and occluded states. (A) Scheme of catalytic cycle of DHFR that shows the two key conformations adopted by the enzyme. The Met20 loop closes the active loop in the E:NADPH:DHF complex, while it is occluded in the E:$NADP^+$:THF complex. (B) Structure of the closed state (CS) (PDB code 1RX2) is on the left and the occluded state (OS) (PDB code 1RX7) is on the right. For clarity, we have explicitly labeled the structural elements that facilitate the allosteric transitions. The major changes are localized in the Met20 loop.

structural transitions may be encoded in the protein structure [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24]. While it is difficult to unambiguously demonstrate whether collective dynamics involving a network of residues facilitates catalysis [7], several experiments and computer simulations suggest that motions of residues that are distal to the active site are important [6, 16, 17]. In the best studied example of *E. coli* dihydrofolate reductase (DHFR) the role of the conformational motions in the enzyme in facilitating the hydride transfer has been illustrated using mutational studies [59, 18, 60], NMR relaxation dispersion measurements [56, 57, 58] that probe the dynamics on μs to ms time scale, molecular dynamics simulations [21, 14, 15], and sequence analysis [14].

DHFR catalyzes the reduction of 7,8 dihydrofolate (DHF) to 5,6,7,8 tetrahydrofolate (THF) [13]. By binding the co-factor, nicotinamide adenine dinucleotide phosphate (NADPH), hydride transfer from NADPH to protonated DHF leads to production of $NADP^+$ and THF [62]. DHFR, which is required for normal folate metabolism in prokaryotes and eukaryotes, plays an important role in cell growth and proliferation in prokaryotes and eukaryotes [63]. As the result of its obvious clinical importance, it has been studied extensively using a wide range of experimental and theoretical methods [13].

High resolution crystal structures show that the *E. Coli* DHFR enzyme has eight β-strands and four α-helices interspersed with flexible loops that connect the secondary structural elements [64, 65]. The structure of DHFR can be partitioned into adenosine binding and loop subdomains (Fig. 2.1 [65]). In the catalytic cycle, Met20 loop changes conformation between closed (CS) and occluded (OS) states (Fig. 2.1). Interactions through hydrogen bond network with the βF-βG loop (residues 117-131) stabilize the CS [66].

The crystal structures of E.*coli* DHFR complexes in the catalytic cycle have given a detailed map of the structural changes that occur in the enzyme [65]. In addition, the conformational changes in E.*coli* DHFR in response to binding have been inferred using various experimental techniques, including X-ray crystallography, fluorescence, nuclear magnetic resonance (NMR) [64, 65, 57]. Comparison of the CS and OS structures shows that the conformations of Met20 loop undergoes the largest change during the reaction cycle. As a result, the states of DHFR are classified using the conformations of the Met20 loop. The active site is either closed,

or occluded depending on the conformation of the Met20 loop (Fig. 2.1). Thus, the motion of the Met20 loop coordinates the dynamical changes in DHFR during the different stages of the catalytic cycle.

More generally, the structural changes that occur in proteins during the reaction cycle, can be thought of as allosteric transitions which are nominally associated with large multi-domain proteins and molecular machines [11, 67]. Allosteric communication between residues that are distal both in sequence and structure is important in the regulation of enzymatic reactions. Starting from the well studied case of oxygen binding to hemoglobin [11], large scale motions have been implicated in the function of a number of biomolecular systems [68]. More recently, even in single domain protein families, allosteric communication is found between functional sites that are spatially apart [26, 69, 70, 71, 45]. Here, we adopt the allosteric perspective to study the kinetics of CS→OS and OS→CS transitions.

Although the structures of the CS and OS states are known the dynamic pathways connecting the two allosteric states have not been characterized [13, 3]. In this chapter, we address the following questions: (1) Can the evolutionary footprints in the DHFR family of sequences be used to obtain a network of residues in DHFR that is linked to the CS→OS and the reverse transition in the enzyme? If so, what role do these residues play in the kinetics of CS→OS and the reverse transitions? (b) What are the pathways and the nature of the kinetics associated with transition from OS to CS and back? (c) What are the structures of the transition state ensemble in the OS→CS transition and in the reverse reaction? We use a combination of bioinformatics methods [25, 45], and Brownian dynamics simulations of

coarse-grained models of DHFR to address these questions. It should be emphasized that our study focuses only on the kinetics of CS→OS and OS→CS transitions, and not on whether the motions that drive these transitions affect hydride transfer reactions. The precise linkage between equilibrium or dynamics motions of proteins and catalysis continues to be a topic of debate [6, 7].

In order to determine the network of residues in DHFR that regulates the allosteric transitions we adopt a sequence-based method [45], which is based on the Statistical Coupling Analysis (SCA) [26, 69, 70, 71]. The SCA identifies many residues that are dispersed between the two subdomains as being relevant in the function of DHFR. Although several of these residues are not strongly conserved, they are predicted to covary across the DHFR family. In order to probe allostery in DHFR, we carried out simulations using coarse-grained Self-organized polymer (SOP) model [50]. The Brownian dynamics simulations reveal the dynamical changes that occur during the CS→OS and OS→CS transitions. The conformational changes in the Met20 loop, which occur by a sliding motion along a helix in the adenosine binding domain, is preceded by coordinated rupture of interactions between Met20 and βF-βG loops and the formation of contacts between Met20 and βG-βH loops. Simulations in which Met16 and Gly121 are crosslinked by a disulfide bond, show that the CS→OS transition is dramatically affected. In accord with the recent NMR experiments [3, 72], we find a small ($\sim (1-3)\%$) of OS (CS) structures are populated by thermal fluctuations when DHFR is in the CS (OS) state. The structures of the transition state ensemble (TSE) is broad both in the forward and reverse direction. The presence of broad TSE and small barrier separating the CS and OS states sup-

ports the conformational selection model that posits that due to the heterogeneous nature of fluctuations conformations resembling the OS state are present in the CS and vice versa.

2.3 Allostery wiring diagram (AWD)

2.3.1 Key residues predicted by SCA are dispersed throughout the structure

We obtained 526 sequences for the DHFR family from Pfam [73] (entry 00186), and realigned them using the Clustalw package [74]. We manually deleted certain sequences, and generated a multiple sequence alignment (MSA) that contained 462 sequences. Each of the 462 sequences has 323 residues including gaps. With the fraction of the sequences in the subalignment set to f=0.35 (see Methods) in the SCA, there are 74 allowed perturbations ($S_j = 0$ for $j = 1, 2, 3...74$) at the various positions in the DHFR family. We used the clustering protocol [47, 75] to identify the set of co-varying residues. After rescaling the $\Delta\Delta G_{ij}$ matrix (Eq. 1.2 in Methods) ($i = 1, 2, 3, ...158$ and $j = 1, 2, 3, ...74$), and using the Euclidean similarity measure in the coupled two-way clustering algorithm [45] we obtained a cluster of 21 residues and a cluster of 19 perturbations. As in our previous work [45], we propose that the residues that are clustered both in positions and perturbations constitute the minimal robust network of residues that signal the kinetics of the CS→ OS transition and back. The relevant network of spatially separated residues constitutes an AWD (Fig. 2.2A), and may encode for the promoting motions.

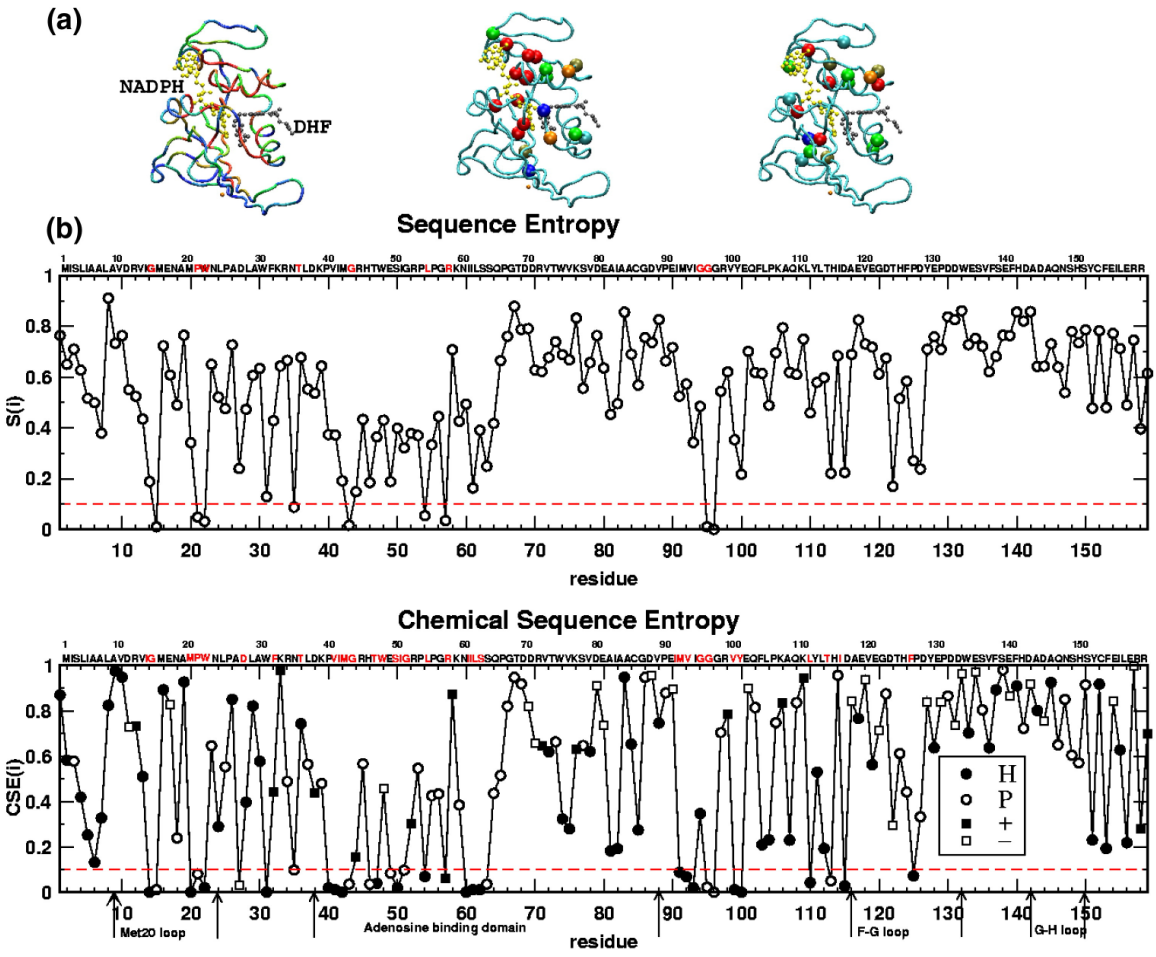


Figure 2.2: Sequence conservation with respect to the structure of the CS. (A) In the CS structure on the left we have color coded the backbone to reflect the extent of sequence conservation. Red color represents strong conservation ($S(i) < 0.1$) and non-conserved residues are in blue. The residues that are clustered in the position in the SCA (V13, I14, N18, M20, D27, L28, K38, V40, I41, M42, W47, I50, G51, N59, I60, L62, Q65, I94, V99, Y111 and T113) are shown in the middle structure. The colors of the residues indicate the values of the $S_{CSE}(i)$ with red representing strong conservation of the chemical identity. The right structure shows the network of residues that appear as perturbation in the SCA (D11, R12, V13, D27, A29, K38, V40, W47, I50, G51, P53, L62, V72, S77, A84, G97, V99, Q108, and Y111). In the three structures the cofactor (NADPH and DHF) are shown using all-atom representation. (B) The top panel shows position dependent sequence entropy ($S(i)$) obtained by aligning the *E. coli* DHFR against the rest of the 426 sequences. Strongly conservation ($S(i) < 0.1$) is observed only for a small fraction of residues. The chemical sequence entropy, $S_{CSE}(i)$ in the bottom panel, shows that for a substantial fraction of residues only the chemical identity changes. These residues are dispersed throughout the structure.

To determine if the residues in the network predicted by SCA merely reflect sequence conservation we calculated the sequence entropy $S_i = -\sum_{x=1}^{20} p_i^x ln(p_i^x)$. For a perfectly conserved residue, $S_i = 0$. If we assume that a residue is strongly conserved if $S_i \leq 0.1$, then there are very few residues with high sequence conservation. These are G15, P21, W22, T35, G43, L54, R57, G95, and G96 (Fig. 2.2B). Sequence entropy is too restrictive in assessing the nature of mutations that are tolerated at a given position. The allowed variations in the amino acid substitution is better captured using the chemical sequence entropy [76], $S_{CSE} = -\sum_{x=1}^{4} p_i^x ln(p_i^x)$ where the twenty amino acids are divided into four classes, namely, Hydrophobic (H), Polar (P), positively charged (+), and negatively charged (-) [76]. Using chemical sequence entropy, residues, namely, I14, G15, M20, P21, W22, D27, F31, T35, V40, I41, M42, G43, T46, W47, S49, I50, G51, L54, R57, I60, I61, L62, S63, I91, M92, V93, G95, G96, V99, Y100, L110, T113, I115, and F125, are strongly ($S_{CSE} \leq 0.1$) conserved.

It is not surprising that many of the residues identified in the AWD (Fig. 2.2) are also strongly conserved as they are associated directly with the binding surface that stabilize the closed conformation. The SCA also identifies residues N18, L28, K38, V72, S77, A84, G97, and Q108 that are neither highly conserved nor adjacent to conserved residues. It appears that many of the residues in the network are relevant for executing dynamical motions that drive the allosteric transitions in DHFR or for cofactor binding. For example, N18 forms contact with H124 in the CS. Similarly, during the OS→CS transition, L28 comes close to I50 upon binding of various ligands which results in the closure of the active site cleft [77]. Residue K38,

which is in the hinge region, facilitates rotation of the adenosine binding domain towards the loop domain (residues M1-D37 and A107-R159) [13].

2.3.2 AWD obtained using SPM overlap with that obtained from SCA

We perform NMA using CS structure, and calculate the overlap between the 100 lowest frequency eigen-modes to the conformational transitions between CS and OS state structures. Minimally three modes, 14, 19 and 28, are found to have largest values of overlap, and account for the transitions between CS and OS. We perform SPM (see chapter 1 for details) for these modes, and select hotspot residues based on the criteria that $\delta\omega = 2 < \delta\omega >$ (see Fig. 2.3). For mode 14, residues in the AWD are within GH-loop, for mode 19, they are located within the Met20 loop, adenosine-binding loop, and GH-loop, and for mode 28, residues in the AWD are within Met20, and FG-loop. Each of the three identified modes encodes for distinct set of residues with large $\delta\omega$ (Fig. 2.3). The elastic energy stored in DHFR are dispersed throughout the structure.

The key residues in the AWD have been shown in previous theoretical and experimental studies to be important either in catalysis or in binding of cofactors. Benkovic and coworkers showed that mutations of residues (M42 and G121) that are far from the active site affect the hydride transfer rates [59, 78]. Based on equilibrium covariance matrix fluctuations of the C_α atoms obtained from all atom MD simulations, Rod et al showed that interactions of M42 with other residues

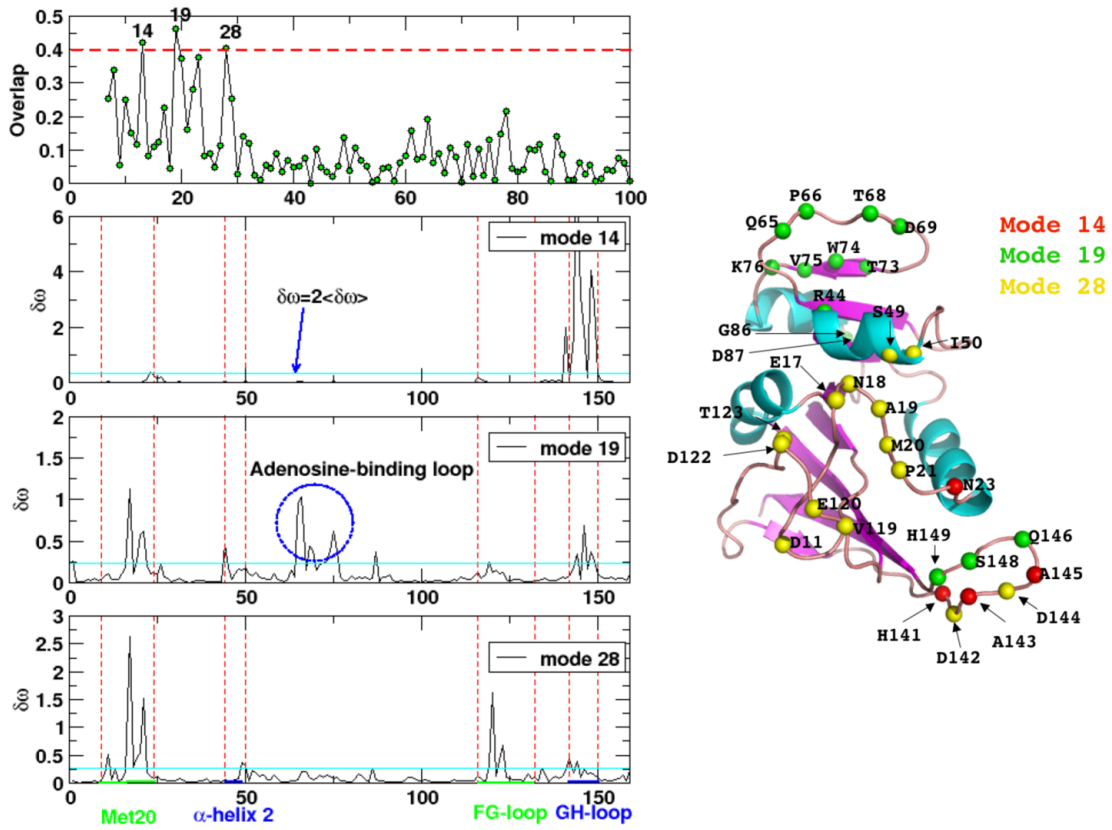


Figure 2.3: Allostery wiring diagram (AWD) of DHFR obtained using SPM. (a) Overlap of the lowest 100 modes with the conformational changes in the the CS→OS transition. Three modes 14, 19, and 28 are identified to have the largest values of overlap. (b) $\delta\omega$ as a function of residue numbers for mode 14. The criteria $\delta\omega = 2 < \delta\omega >$ is used to identify hotspot residues involved in AWD. (c) Same as (b) except this is for mode 19. (c) Same as (b) except this is for mode 28. On the bottom of (c), the division of important structural units, Met20 loop, $\alpha-$helix 2, FG-loop, GH-loop, are indicated and labeled. (d) The hotspot residues identified from mode 14, model 19, and mode 28 are indicated with red, green, and yellow spheres, respectively. Met20-loop is identified in both modes 19 and 28, and is shown with yellow spheres.

(H45, D28, S49) would also be involved in the CS→OS conformational transition [21]. Mutations of positions M42 and/or G121, that lead to anti-correlated motions between the two subdomains, are found to be part of the predicted AWD. Hammes-Schiffer used sequence conservation of a small dataset of DHFR sequences to identify

a network of residues whose coordinated motion is apparently linked to catalysis [14, 15]. Among them, I14 is found to be in the AWD that we have identified using SCA. It was also found that motions of residues W22, D27, M42, I60, L62, and T113 which forms hydrogen bond network with DHF in the active site might also be involved in coupled promoting motions [14, 15]. Taken together, the present and previous studies show that the AWD, that represents the network of signaling residues in DHFR, is spread throughout the structure (Fig. 2.2). More importantly, many non-conserved residues are part of the network. In an recent study, Mauldin *et al* used NMR to characterize DHFR in complex with drugs. They found inhibitors simultaneously modulate the closed and occluded transitions, as well as the fast motions at distal regions of DHFR. The residues they identified to link the dynamics and catalysis of DHFR agree well with our predictions [79].

2.4 Kinetics analysis of DHFR signaling pathways

2.4.1 Anticorrelated Motions of DHFR in the CS and OS states

The Root Mean Square Deviation (RMSD) between the closed and occluded crystal structures of *E. Coli.* DHFR is only 1.18 $\mathring{A}$. However, the RMSD of the active Met20 loop (residues A9-L24) between the two end point structures is alomost three times larger ($\approx$ 3.35 $\mathring{A}$). In order to assess the differences in the structures of the two states at finite temperature, we equilibrated the OS and CS conformations at 300 K. Comparison of the thermally averaged contact maps shows that the closed state differs from the occluded conformation mainly in the Met20 loop and the

secondary structural elements that are affected by the motions (see below) of the Met20 loop (data not shown). The largest changes occur in the βF-βG (D116-D132) and βG-βH loop (D142-S150) loops, and the α-helix H2 of the adenosine binding domain(residues R44-I50). The crystal structures and the thermally equilibrated CS and OS states also show that, in the CS→OS transition the conformational fluctuations in the Met20 loop have to be accompanied by the following changes: (1) Contacts between the Met20 and βF-βG loops should be ruptured. (2) Interactions between helix 2 and the Met20 loop should be disrupted, and reform in a different location; (3) Stabilizing contacts between Met20 and βG-βH loops should form. If these processes are disrupted then it is likely that the catalytic efficiency of DHFR may be compromised. Indeed, experimental findings of the importance of mutating M42, G121, S148 or any two of these residues on the hydride transfer rates can be rationalized based solely on a static picture [21].

The correlated motions in DHFR are computed using time average covariance matrix defined as,

$$\langle C_{ij}(X)\rangle = \frac{1}{T_{obs}} \int_0^{T_{obs}} \Delta\hat{r}_i(t) \cdot \Delta\hat{r}_j(t) dt \tag{2.1}$$

where $\Delta\hat{r}_i(t) = (\vec{r}_i(t) - \vec{r}_i^{wt})/|\vec{r}_i(t) - \vec{r}_i^{wt}|$ is the unit vector of the displacement of the i^{th} C_α atom with respect to its initial value, and X is either CS or OS. The direction of motion of the i^{th} residue is given by $\Delta\hat{r}_i$. If $\langle C_{ij}\rangle$ is positive then the motion of the two residues i and j are correlated while negative values correspond to anti-correlation. For perfectly correlated (anti-correlated) residues

$\langle C_{ij} \rangle$ is +1 (-1). The covariance map for the CS shows anti-correlated motion between the Met20 loop and the adenosine binding domain, as well as between the βF-βG and βG-βH loops (see the dark blue regions in Fig. 2.4). Thus, the adenosine binding domain and the loop domains move in an anti-correlated manner. Similar conclusions were obtained using all atom simulations of the WT DHFR [21]. The present simulations and previous MD studies [21, 23, 14, 15, 16, 17] point to the importance of correlated motions between regions that are spatially well separated. The cross correlations in the inter-domain motions shown in Fig. 2.4 are obtained by averaging the structural fluctuations over 0.1 ms, and may well be relevant in facilitating cofactor binding and solvent rearrangement needed for catalysis [80]. The static picture alone is not sufficient to describe the kinetics of transitions between the CS and OS. Only by probing the kinetics of conformational fluctuations in the CS→OS (and the backward) transitions we can predict the order of events that results in the conformational changes in the all important Met20 loop.

The bottom panel in Fig. 2.4 shows the differences in the covariance matrices $\Delta C_{ij} = \langle C_{ij}(CS) \rangle - \langle C_{ij}(OS) \rangle$ in regions that are significantly different between the two allosteric states. The red region between the Met20 loop and the adenosine binding domain indicates more anti-correlated motions in OS than in CS. The blue region between the Met20 loop and the other two loops shows less anti-correlated motions and more correlated motions in OS than in CS.

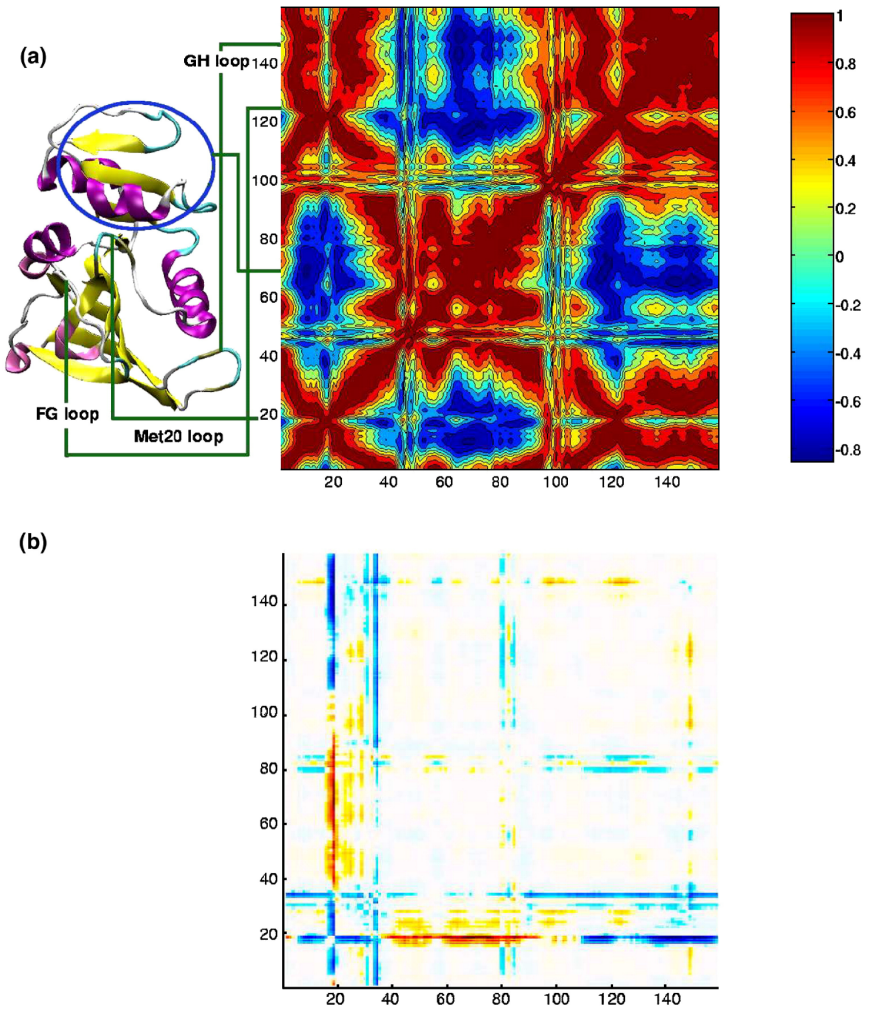


Figure 2.4: Correlated and anti-correlated motions in DHFR. Covariance matrix of equilibrium fluctuations of the unit vectors constructed from the coordinates of the C_α atoms (Eq. 2.1) of the wild type DHFR in the CS. The residues associated with the structural elements are shown on the left. The scale on the right measures the extent of correlation. In the bottom panel we show $|C_{ij}^{CS} - C_{ij}^{OS}|$ using the same scale. For clarity, we only highlight those regions that are different in the two allosteric states. The simplicity of the SOP model has allowed us to probe the equilibrium fluctuations on about 0.1 msec.

2.4.2 Deformation of the Met20 loop in CS→OS transition

In order to dissect the kinetics of structural changes in the Met20 loop during the forward (CS→OS) and the backward (CS→OS) directions we have performed

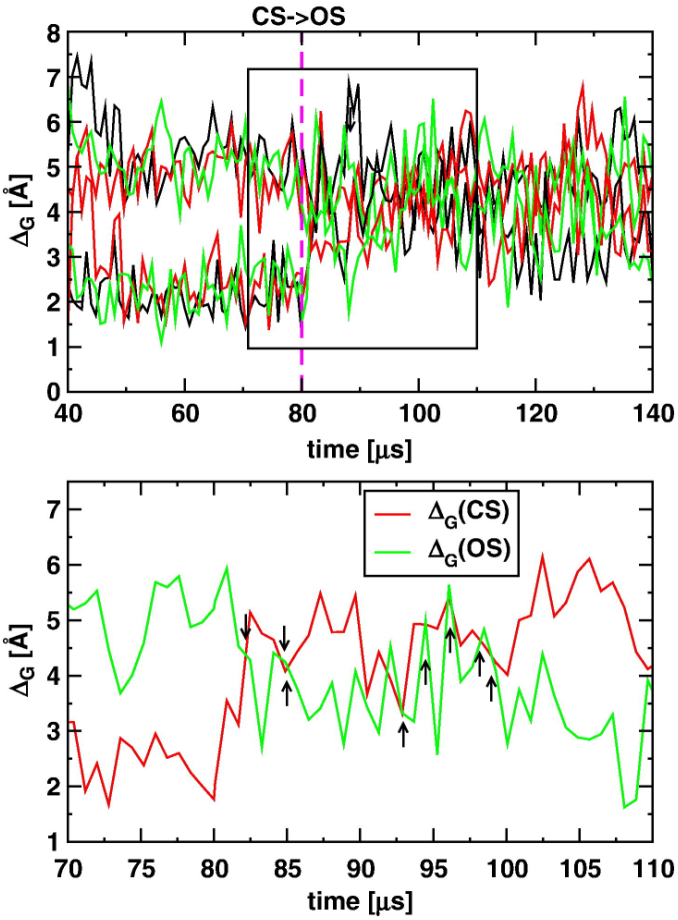


Figure 2.5: RMSD as a function of time during the CS→OS transition. Time dependent changes in the global RMSD ($\Delta_G(t)$) for a few representative trajectories as a function of t are given in the top panel. The dynamical changes in $\Delta_G(t)$ for one of the trajectories for the time interval enclosed by the box are shown below. The arrows show that frequent recrossings between CS and OS states occur prior to the completion of CS→OS transition.

a number of simulations using the procedures described in the Methods section. Although it is clear that Met20 loop plays essential role in this transition, the order of events that drives its conformational change is not known [13, 3]. We monitor the Met20 loop kinetics using two surrogate reaction coordinates. One is the global RMSD, Δ_G, that is obtained by aligning the instantaneous conformation of DHFR at time t either with the CS or the OS structure. During the CS→OS transition,

Δ_G, should increase with respect to the CS and decrease with respect to OS. From the time dependent variations in Δ_G we can infer the changes in the Met20 loop with respect to the entire structure. To determine the changes that are localized in the Met20 loop we calculated a local RMSD, Δ_L, which uses only the coordinates of the active loop. From the time dependent changes in Δ_L, which is computed by aligning the Met20 loop and computing its RMSD (with respect to the starting conformation) during the two transitions, we can explicitly identify the dominant motions (translation, rotation, or twist) of the loop. The kinetics expressed in terms of the local coordinate Δ_L yields the conformational changes of only the Met20 loop.

The time dependent changes in the global RMSD, $\Delta_G(t)$, with respect to the CS show considerable dynamical heterogeneity (Fig. 2.5). The bottom panel in Fig. 2.5, for one trajectory, shows that there are multiple recrossings across the transition region which is suggestive of a rather broad transition region (see below). Prior to the CS $\rightarrow$ OS transition ($t < 80\mu s$ in Fig. 2.5) $\Delta_G(t)$ undergoes substantial fluctuations which suggests that high energy states are being sampled while DHFR is in the CS. More importantly, such fluctuations can lead to infrequent visits to conformations that are similar to the OS state (see below). The broad distribution of transition times and multiple recrossings attests to the plasticity of the enzyme during the conformational transition.

Although there is great diversity in the dynamics of the individual trajectories $< \Delta_G(t) >$ and $< \Delta_L(t) >$, obtained by averaging over an ensemble of initial conformations, can be approximately described using a two-state model (Fig. 2.6 A). Comparison of $\Delta_G(t)$ and $\Delta_L(t)$ shows (Fig. 2.6A) that deformations of the

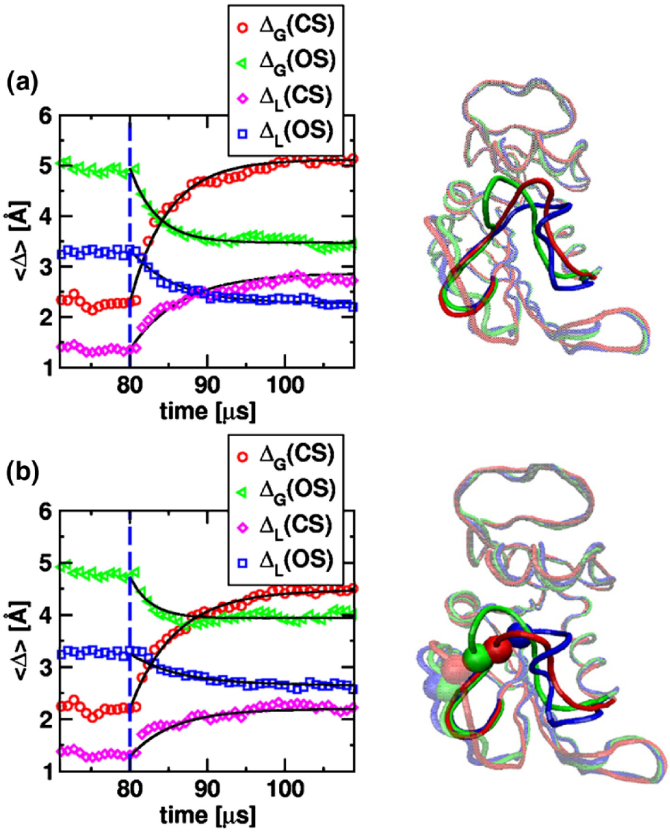


Figure 2.6: Kinetics of allosteric transitions probed using RMSD. Time dependent changes in the global, $< \Delta_G(t) >$, and the local, $< \Delta_L(t) >$, RMSD of the Met20 loop averaged over 50 trajectories. The meaning of the symbols are given in the insets. The RMSDs are measured with respect to the starting and ending states. For example, $< \Delta_G(CS) >$ means that the global RMSD is computed with respect to the CS. The changes in $< \Delta_G(t) >$ and $< \Delta_L(t) >$ for the WT CS→OS are shown in (A), and (B) shows the results for the CL. The structures on the right in both (A) and (B) represent superposition of the CS, OS and the average transition state (TS) conformations. Conformation of the Met20 loop in the CS (green), OS (blue), and TS (red) are highlighted. The cross link between 16 and 121 is explicitly shown in (B).

Met20 loop occurs after the global motions in the CS→OS transition. Because the long time values of $< \Delta_L >$ are less than $< \Delta_G >$ values, we surmise that the structural changes in the Met20 loop involve translation and rotational motion towards the OS structure.

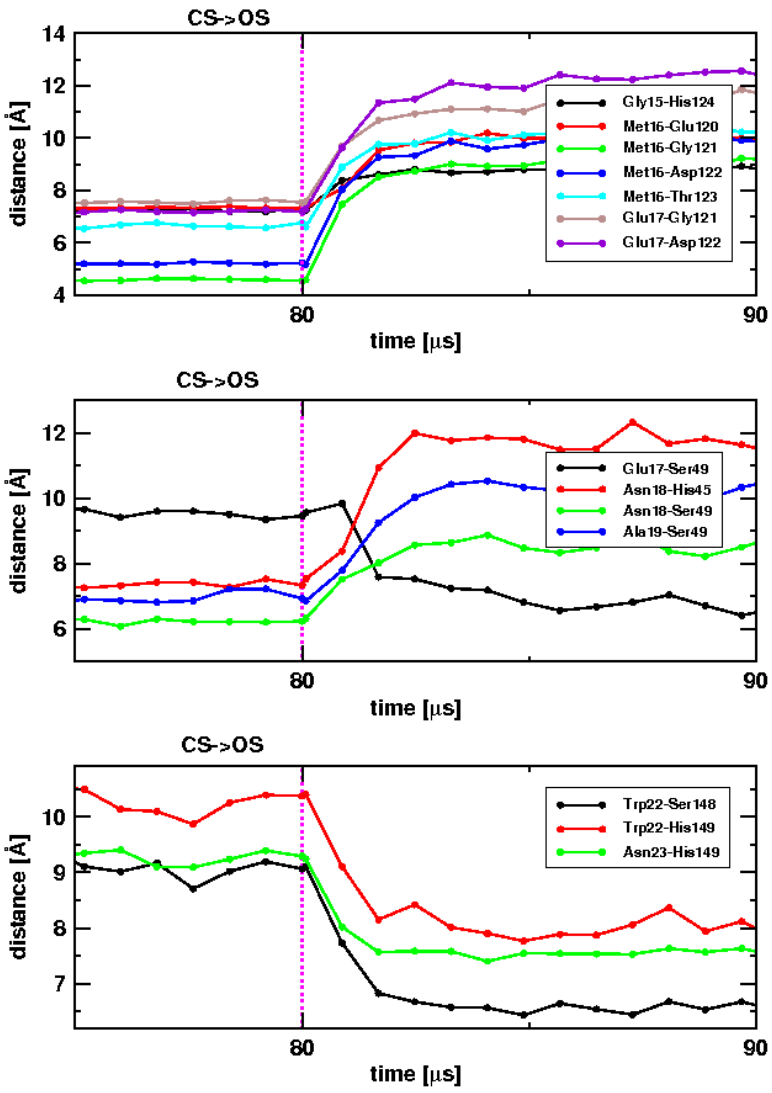


Figure 2.7: Kinetics of rupture of individual contacts between residues in the various substructures of DHFR during the CS→OS transition. (A) This panel shows the changes in the distance between the residues in the Met20 loop and βF-βG loop that rupture and form. (B) Same as (A) except the residues represent interactions between the Met20 loop and α2. (C) shows the kinetics describing the formation of interactions between the Met20 loop and βG-βH loop. In all cases the identifies of the residues are shown on the right.

2.4.3 Sliding of the Met20 loop across α2 limits the CS→OS transition rate

In order to understand the mechanism of the communication during the CS→OS transition we monitored the local movements of the Met20 loop and the helix α2

in the adenosine binding domain. Rupture of the contacts in the CS state (Asn18-His45, Asn18-Ser49, and Ala19-Ser49) and formation of Glu17-Ser49 during the CS→OS transition facilitates the sliding of Met20 along α2 (Fig. 2.7A). The relative sliding motion between α2 and the Met20 loop enables NADPH to move closer to DHF.

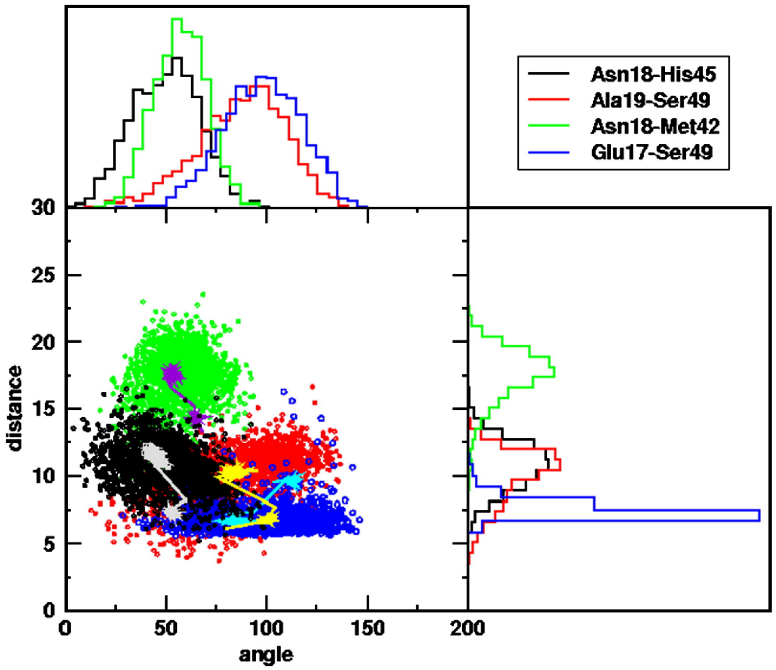


Figure 2.8: Illustration of the sliding of the Met20 loop along α2. Two dimensional projection of the distance (R_i) and angle (α_i) that are kinetically sampled in the 50 trajectories during the CS→OS transition. The angles α_i are defined in the text. The monitored residues are identified on the upper right corner. The colors grey, yellow, purple and cyan represent ensemble averages. The monotonic decrease in all the angles shows sliding of the Met20 loop along α2. Histograms of the distance and the angles for the residue pairs are displayed on the right and on top respectively.

In the loop subdomain, the flexible Met20 loop interacts simultaneously with both the βF-βG and βG-βH loops 2.1. In order dissect the order of events that occurs in the CS→OS transition we have computed the kinetics of breakage and formation of a number of contacts involving the two loops (Fig. 2.7A-C). By

fitting the time-dependent changes in the formation and rupture of contacts to single exponential kinetics we find that the rupture of contacts between Met20 loop and βF-βG loop in CS as well as formation of contacts between residues in the Met20 loop and βG-βH loop occur nearly simultaneously (on the μs time scale) (see (Fig. 2.7A-C)). Only subsequently (on a time scale of about 2 μs), the interaction between Glu17 (in Met20 loop) and Ser49 (in α2) that exists only the CS state, takes place. Thus, the sliding of Met20 loop on α2 requires coordinated motion of a number of residues in the loop domain.

We can further dissect the nature of the sliding motion of the Met20 loop along α2 by simultaneously measuring the changes in the angles and the distances between selected residues. We have computed the time-dependent changes in the distances between Asn18-His45 (R_1), Ala19-Ser49 (R_2), Asn18-Met42 (R_3) and Glu17-Ser49 (R_4), respectively. The sliding motion is vividly illustrated using the changes in the angles that the vectors $\vec{R}_1$, $\vec{R}_2$, $\vec{R}_3$ and $\vec{R}_4$ make with the axis of α2. Angles are defined as $\alpha_i = cos^{-1}(\hat{R}_i \cdot \hat{U}_{H2})$ $i = 1, 2, 3, 4$ and $\hat{U}_{H2}$ is the unit vector of the α2 helix axis. The two-dimensional projection of (R_i, α_i) $(i = 1, 2, 3, 4)$, that represents the values of (R_i, α_i) that are sampled in the kinetic trajectories, shows that α_i values decrease monotonically during the CS$\rightarrow$OS transition. The averages over all the trajectories for α_i also show a monotonic decrease. The averages also show that the α_i values are either clustered around the CS or the OS state. In other words, there is very little backtracking in the sliding movement of the Met20 loop along H2. The histogram of the angels and distances sampled during the transition in Fig. 2.8 also shows the fluctuations in α_i $(i = 1, 2, 3, 4)$ are centered around the OS values

which suggests that (in terms of these microscopic variables) that the transition occurs when the conformation is close to the OS state. This result is also in accord with the results in Fig. 2.7A which shows that Glu17-Ser49 only forms when the interactions in the CS are ruptured, a process that occurs closer to the completion of the CS → OS transition. The structural changes that accompany the sliding motion of the Met20 loop involves concerted motion of a number of residues (see the diagram on the left in Fig. 2.9). The figures summarize the collective motions of residues in both the subdomains that facilitate the structural deformations in the Met20 loop.

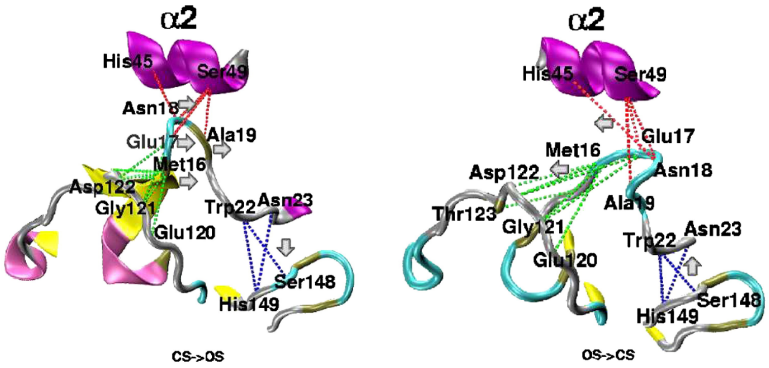


Figure 2.9: Structural representation of the coordinated changes in the distances of residues that accompany the sliding motion in the CS→OS (left side) and OS→CS (right side) transition. The arrows indicate the direction in which the structural changes occur. The displayed structural changes were inferred from the kinetics shown in Figs. 2.7, 2.12, and 2.8. In the forward transition the Met20 loop is pulled by the βG-βH loop which results in it being pushed away from βF-βG loop. The push-pull process results in the sliding of the Met20 loop. The mechanism is approximately reserved in the OS→CS transition.

2.4.4 Cysteine crosslink inhibits CS→OS transition

The kinetics in both the forward and the backward (see below) transitions show that the coordinated motion in the loop subdomain plays an important role in enabling the Met20 loop communicate with adenosine binding domain. In the crystal structure of CS, the distance between the C_α atoms of Met16 and Gly121 is about 4.3$\mathring{A}$. It is possible to mutate these residues to Cys to establish a disulfide cross link. We have simulated the kinetics of the CS→OS transition in the crosslink mutant (referred to as CL) to assess the extent to which the motion of the Met20 loop is inhibited. Previously, it has been argued that constraining even residues that are 28 $\mathring{A}$ apart can affect hydride transfer rates [81]. Our purpose in studying the CL mutant is to see how the strain in the loop domain would affect the communication between the two domains. Since the disulfide bond constrains the distance between Met16 and Gly121 to 4.3 $\mathring{A}$, the anti-correlated motion between Met20 loop and βF-βG loop should be impeded. The time dependencies of $< \Delta_L(CS|t) >$ and $< \Delta_L(OS|t) >$ show that the Met20 loop does not fully adopt its conformation in the OS state (compare Fig. 2.6A and B). Similarly, the long time values of $< \Delta_G(CS) >$ and $< \Delta_G(OS) >$ in the mutant are different than in the WT (see Fig. 2.6B). In the WT, βG-βH loop are involved in the coordinated motion between two domains. Surprisingly, the crosslink has little effect on the relative motion between βG-βH loop and the Met20 loop. The time dependent changes that monitor the formation of contacts between Trp22-His149 and Asn23-His149 are similar in the WT and the crosslink mutant. Because the interactions between the Met20 loop and βG-βH loop

are not fully inhibited in the CL, the sliding motion across $\alpha 2$ with the formation of Glu17-Ser49 can occur (Fig. 2.12A) albeit less efficiently. We predict that due to the incomplete CS→OS transition the crosslink will dramatically affect the rate of the forward hydride transition. Experiments using CL can shed further light on the importance of enzyme motion in catalysis which still remains controversial [14, 15, 7].

2.4.5 Deformation of the Met20 loop in the OS→CS transition

The time constant for the local kinetics of the Met20 loop in the OS→CS transition obtained from $< \Delta_L(t) >$ (Fig. 2.10) is greater than the time scale in which $< \Delta_G(t) >$ changes. This implies that the Met20 gliding across $\alpha 2$ is the first event in the OS→CS transition. In contrast, during the CS→OS transition, only in the final stages does the Met20 loop occludes the active site.

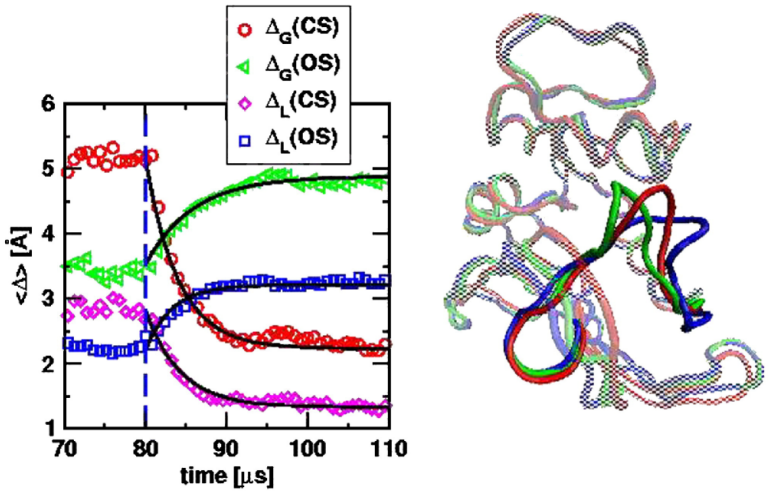


Figure 2.10: Changes in $\Delta_G(t)$ and $\Delta_L(t)$ for the OS→CS transitions. Same as Fig. 2.6 except that the transition is from OS→CS. The conformation of the Met20 loop in the TS (red) is different from that in Fig. 2.6.

Although the initial change in the OS→CS transition involves the deformation of the Met20 loop (Fig. 2.10) the microscopic events that drive this transition are distinct from those seen in the CS→OS transition. Remarkably, the rupture of Glu17-Ser49 occurs only after the formation of the contact between the Met20 loop and α2. The time dependent changes in the contacts present only in the CS state (Asn18-His45, Asn18-Ser49, and Ala19-Ser49) occur while Glu17-Ser49 (in the OS state) contact still persists (Fig. 2.12B). We suggest that binding of NADPH, which is required for THF to be released, assists in the formation of contacts between Met20 loop and βF-βG, and between Met20 loop and α2. Only after these contacts are established the contact between Glu17-Ser49 ruptures (Fig. 2.12B). Upon rupture of the Glu17-Ser49 contact, the Met20 loop slides back to its closed conformation, and THF is released.

The simulations also show coordinated motions among the three loops in the loop subdomain during the OS→CS transition (see the right side of Fig. 2.9). From the analysis of the time-dependent changes in the distances between a number of residues we conclude that βF-βG loop stretches the Met20 loop by forming a number of contacts (Gly15-His124 ,Met16-Glu120, Met16-Gly121, Met16-Asp122, Met16-Thr123, Glu17-Gly121, and Glu17-Asp122) with the Met20 loop. In concert with these events the strain imposed on the Met20 loop by the βG-βH loop is released by rupture of contacts (Trp22-Ser148, Trp22-His149, and Asn23-His149) with the Met20 loop. The pull (by the βF-βG loop) and push (by the βG-βH loop) action on the Met20 loop must take place before the Met20 loop slides back to its conformation on the CS state (Fig. 2.7B). These results show that the pathways in the OS→CS

transition are not the reverse of what transpires during the CS→OS transition. The structural changes in the Met20 loop and the concerted motions of a number of residues that drive these changes are shown on the right side of Fig. 2.9.

2.4.6 Transition state emsemble

We have used the global RMSD (Δ_G) as a surrogate reaction coordinate to determine the structures of the transition state ensemble (TSE). We assume that the transition state (TS) for a molecule is reached for the first time at t_{TS}, if $|\Delta_G^C S(t_{TS}) - \Delta_G^O S(t_{TS})| < \epsilon (= 0.5\mathring{A})$ is satisfied. Our criterion places the transition state equidistant (in terms of the global RMSD) from the CS and OS. Comparison of the contact maps (data not shown) for the TSE, CS, and OS shows that both the transitions exhibit major changes more with respect to the starting than the ending state. The largest changes between the CS and OS states, which take place in the Met20 and $\beta F - \beta G$ loops, occur before the transition state is reached.

The heterogeneity observed in the dynamics of the CS→OS transition is also reflected in the distribution $P(t_{TS})$ of the transition time t_{TS} (Fig. 2.11A). Surprisingly, $P(t_{TS})$ is approximately uniform in the CS → OS transition (Fig. 2.11A). As a result, the TSE structures are much less heterogeneous in the forward than in the backward direction (Fig. 2.11C). However, the spread in t_{TS} is broader in the forward direction compared to the backward direction. From the TSE we can compute a Tanford β-like parameter, $q^{\ddagger}$, $(0 \leq q^{\ddagger} \leq 1)$ using

$$q^{\ddagger} = \frac{max(\Delta^{\ddagger}) - \Delta^{\ddagger}}{max(\Delta^{\ddagger}) - min(\Delta^{\ddagger})} \tag{2.2}$$

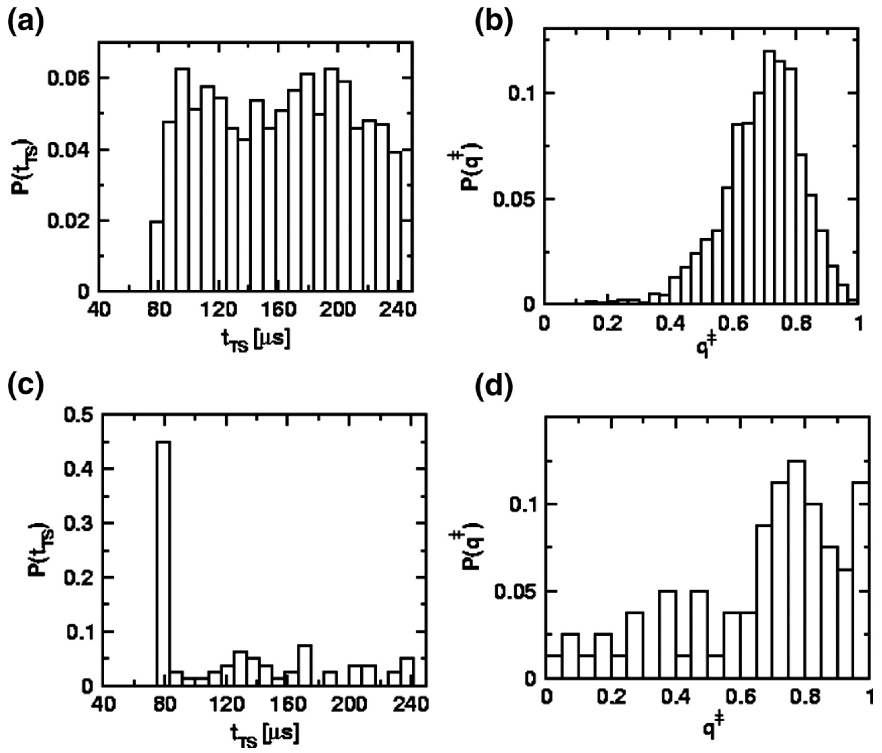


Figure 2.11: Characteristics of the transition state ensemble (TSE). (A) and (C) show the distribution of transition times $P(t_{TS})$ for the forward and reverse transition, respectively. (B) and (D) represent the distributions $P(q^{\ddagger})$ of $q^{\ddagger}$ (see Eq. 2.2) for the CS→OS and OS→CS transitions respectively. In both cases the TSE is broad. However, the width of the TSE (inferred from $P(q^{\ddagger})$) in the reverse direction is larger. The fluctuation $\frac{<q^{\ddagger 2}>-<q^{\ddagger}>^2}{<q^{\ddagger}>^2}$ is 0.2 for CS→OS and 0.6 for OS→CS.

where $\Delta^{\ddagger} = (\Delta_G^C(t_{TS}) + \Delta_G^O(t_{TS}))/2$, $max(\Delta^{\ddagger})$ and $min(\Delta^{\ddagger})$ are the maximum and minimum values of $\Delta^{\ddagger}$ respectively. If $q^{\ddagger}$ is close to 0 (1) then the most probable TS is starting (ending) allosteric state. For the CS →OS transition the average value of $q^{\ddagger}$ is 0.66 (Fig. 2.11B) which implies that the TSE structures are more OS-like (see the average TS structures in Figs. 2.6 and 2.10). Although the distribution $P(q^{\ddagger})$ for the OS→CS is very broad (Fig. 2.11D) the most probable $q^{\ddagger}$ is closer to the CS than to the OS. Thus, in both the transitions the average TSE structures resemble the high energy allosteric states. This observation supports the recent inferences drawn

from the NMR relaxation time measurements [3] that the high energy conformation is populated (in the pre-equilibrium sense) in the both the allosteric states. From the Hammond postulate it follows that the TSE structures should resemble the high free energy states in accord with the present simulation results. Surprisingly, we further predict that the TSE structures for the OS→CS transition is conformationally much more heterogeneous than in the forward direction (Compare Figs. 2.11B and 2.11D).

2.5 Residues in the AWD code for ligand binding and dynamics

The SCA predicts a number of residues that are expected to be relevant either in the motion of DHFR or in the function (Fig. 2.2A). Some of the residues in the network are related to cofactor binding and interaction with the active site while others are directly involved in accommodating the motion of the Met20 loop during the CS→OS transition. For example, SCA identified Leu28, Ala29, and Ser63 (Fig. 2.2) all of which are involved in ligand binding or binding-involved dynamics. The amino acid at location 29, which in *E. Coli* DHFR is Ala, is in contact with His28 show isomerization between two isoforms of the apoenzyme [13]. In *L. casei* enzyme the conversion between the isoforms occurs only for the folate-bound complex while in human DHFR there appears to be only conformation in the methotrexate (MTX) DHFR complex [13]. The importance of Ser63 in maintaining hydrogen bond with NADPH was noted in the molecular dynamics simulations [14, 15]. Similarly, Asp27 is involved in hydrogen network with DHF in the active site [77]. The network predicted by SCA also contains Ile60 and Leu62 both of which have been recognized

to be dynamically involved in interactions with Met20 loop. SCA also suggests that Ile94 and Gly97 should play a role in the function of DHFR. Because SCA cannot assess the importance of absolutely conserved residues it is likely that neighboring residues Gly95 and Gly96 may be relevant in the reaction cycle of DHFR [14, 15]. It is noteworthy that the SCA identified a network of residues in the helical region $\alpha 2$ in the adenosine binding domain as being important. The present simulations show that the critical sliding motion of the Met20 loop along $\alpha 2$ completes the allosteric transitions. Mutations in the region (Ile41-His45), that is far from the active site, have great influence on the forward hydride transfer reaction without affecting cofactor binding [16, 17]. It appears that the predictions of the SCA can be rationalized in light of a number of experimental and theoretical studies that have identified the importance of concerted motions among a sparse network of residues on the reaction cycle of DHFR. The sequence-based approach fails to identify key residues (Gly121 being the most important) which apparently plays a role in catalysis [59].

2.6 A small fraction of OS (CS) is present under equilibrium conditions in the CS (OS) state

Although the dominant basin of attraction corresponds to a unique native folded state enzymes can sample other conformations, albeit not frequently, through thermal fluctuations. Some of the conformations that are sampled in the ensemble of the equilibrated CS can correspond to the structures in the OS [3]. Allosteric

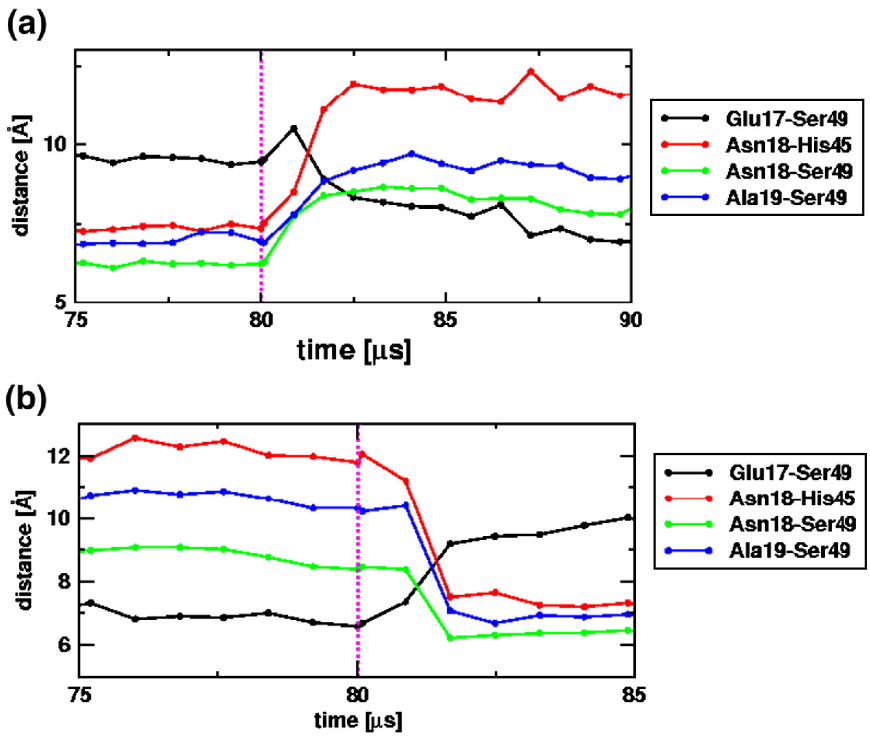


Figure 2.12: Dissection of the local changes in the kinetics in CL (CS→OS) and WT (OS→CS). (A) Time-dependent changes in the distances between select residues from the Met20 loop and $\alpha 2$ for the CS→OS transition in the CL. The transition is from CS→OS. (B) Same as (A) except these represent changes that occur during the OS→CS transition for the WT.

mechanism based on the preexisting equilibrium [82] is qualitatively different from the induced-fit model [83] which posits that the conformational transitions in the CS state occur only after the ligand binds. Indeed, several experiments, including the recent reports on DHFR [3], suggest that a small population of OS conformations are in equilibrium with the CS structures. Similarly, we expect that CS structures should be accessible when the molecule is predominantly in the OS.

In order to probe the validity of the conformational selection model [82] we calculated the distribution $P(\Delta\Delta_G)$ where

$$\Delta\Delta_G = \Delta_G^{OS} - \Delta_G^{CS} \tag{2.3}$$

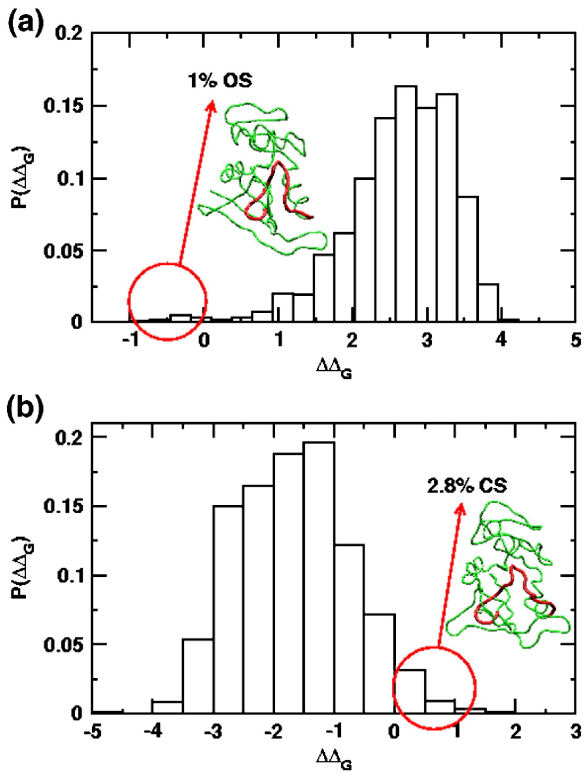


Figure 2.13: Sampling of OS (CS) in CS (OS) state. (A) Distribution of $P(\Delta\Delta_G)$, calculated using an ensemble of equilibrated conformations in the CS, as a function of $\Delta\Delta_G$ (Eq. 2.3). The negative regions represent sampling of conformations that resemble the OS. (B) Same as (A) except $P(\Delta\Delta_G)$ is obtained from an ensemble of equilibrated structures in OS state. Under equilibrium conditions a minor population $\sim (1-3)\%$ of the product-like structures are present. The displayed structure in (A) is OS-like while the one in (B) is CS-like.

where Δ_G^{OS} is the equilibrium RMSD of conformations in the CS with respect to the OS structure, and Δ_G^{CS} is the corresponding RMSD with respect to the CS structure. If DHFR is in the CS without ever sampling the OS-like structures then we expect that $\Delta_G^{CS} \approx 0$. As a result, $P(\Delta\Delta_G)$ should be identically zero whenever $\Delta\Delta_G < 0$. Thus, the observation of negative values of $\Delta\Delta_G$ is an indication of preexisting OS-like structures even under equilibrium conditions that favor CS (Fig. 2.13A). Figure 2.13A (Fig. 2.13B) shows that $P(\Delta\Delta_G)$ is non-zero for a small range of negative

(positive) $\Delta\Delta_G$ in the ensemble of CS (OS). The population of the micro species CS (OS) in the OS (CS) basin is $\sim 3\%$ ($\sim 1\%$). Surprisingly, these estimates are similar to the values reported by Boehr *et al.* [3]. The presence of higher energy species also suggests, in accord with the Hammond postulate that the TS structure should be OS-like in the CS→OS transition. This inference, which follows from the conformational selection model is in accord with our simulations. We predict that mutations that destabilize either the CS or the OS will affect the kinetics of the allosteric transitions.

2.7 Concluding Remarks

From the perspective of allostery, it is not surprising that communication between residues that are spatially well separated facilitates the CS→OS transition [67]. We have used sequence-based method to identify a network of mechanically important residues that could control the kinetics of conformational transitions. The residues in the network are dispersed both in the adenosine binding and the loop domains. Coordinated motions among these residues and others control the structural transitions, and perhaps the forward and backward hydride transfer reaction. Surprisingly, several of these residues are not strongly conserved although their chemical character are often preserved across the various species. Hydride transfer experiments on the wild type DHFR and its mutants [6, 18] have already pointed out the importance of many of the residues in the network. In addition, all atom molecular dynamics simulations [14, 21] and NMR experiments [13, 3] have

implicated the key role of the network residues in the dynamics of DHFR. Although it is difficult to unambiguously establish a direct link between the DHFR motions (equilibrium or dynamic) and hydride transfer reaction [7], the perturbation of these residues will affect magnetic resonance relaxation dispersion.

The kinetics of the allosteric transitions in the forward (CS→OS) and the reverse (OS→CS), using the SOP model, reveal in great detail the order of events that results in the movement of the Met20 loop. In the forward direction, several contacts in the CS state rupture and new ones form in the OS. The concerted kinetics associated with these contacts, most of which are associated with the Met20, βF-βG, and βG-βH loops facilitate the sliding motion of the Met20 loop so that it occludes the active site (Fig. 2.1). Surprisingly, the pathways in the OS→CS transition are not the reverse of the forward reaction. In particular, the interactions between Glu17-Ser49, whose rupture facilitates the sliding of the Met20 back to its CS position, persist till late in the OS→CS transition. In the forward direction, Glu17-Ser49 contact occurs late for the sliding motion of Met20 along α2 to take place. The broad transition state region, both in the forward and backward directions, attests to the inherent plasticity of enzymes in general, and DHFR in particular. These results support the notion that mutations that inhibit the equilibrium fluctuations leading to the population of the minor species can adversely affect the rates of hydride transfer reaction. Indeed, the observed decrease in the hydride transfer rate in G121V has been rationalized using this picture [59].

Of particular importance is the link between the present studies and the recent NMR relaxation measurements Boehr *et al.* [3] which showed that, at equilibrium,

there is a small percentage of OS structures in the ensemble of CS conformations. Similarly, when DHFR is in the OS state dynamical fluctuations populate a small ($\sim$ $(1-3)\%$) of CS structures. Our simulation results are in accord with the NMR experiments [3]. These results support the emerging notion that in enzymes conformations resembling the cofactor-bound structure is already present in the apoenzyme. The cofactors dynamically funnel the minor populations so that the equilibrium shifts to the haloenzyme. The present simulations show that such conformational fluctuations occur on μs time scale. Because of the simplicity of the SOP model the estimated time scale should be taken as a lower bound. The ability to access the higher free energy states on (μs-ms) time scale is a consequence of the conformational heterogeneity of the enzyme which leads to low barriers separating the relevant kinetic states. In DHFR, this is reflected in the broad TSE with heterogeneous structures that results in a broad distribution of crossing times between the allosteric states.

We also obtained the temperature (T) dependence of the rates of the forward and reverse transition for the WT and the forward transition for the CL. The rates were computed by fitting the time dependence of $< \Delta_G(t) >$ for T in the range $285K < T < 315K$. The averaging is performed using 20 trajectories. We find that the three rates follow the Arrhenius behavior. Because the SOP is a coarse-grained model the activation barrier is severely underestimated. Nevertheless, the results show that the rates of the allosteric transitions are enhanced as T increases. To the extent the enzyme motion is linked to the hydride transfer from NADPH to dihydrofolate, we can conclude that such motions must facilitate catalysis [84].

Needless to say that altering T also changes reorganization free energies of the solvent which could be a dominant factor in determining the catalytic rates [7, 80].

The SOP model, which was introduced to carry out simulations of large systems [35, 52], does not include a number of relevant interactions. Most notably, the lack of explicit model for hydrogen bonds, prevents us from examining their role in the allosteric transitions. The role the network of hydrogen bonds of DHFR plays in affecting the CS→OS transition can only be vicariously gleaned using the SOP model. On the other hand, the major advantage of the SOP model is that long time simulations for a large number of trajectories can be carried out. Indeed, the non-trivial prediction that the coordinated motions of specific residues throughout the structure trigger the movement of the Met20 loop is amenable to experimental tests. The non-conserved residues identified in this work can form the basis of future mutagenesis experiments. We believe that a combination of computational methods (sequence-based technique, coarse-grained and all atom MD simulations), and NMR, single molecule, and biochemical experiments are needed to fully dissect the interplay between enzyme motion and catalysis.

2.8 Method

In order to identify the residues that are involved in transmitting allosteric signals, we used our formulation [45] of the SCA introduced by Lockless and Ranganathan [26, 69, 70]. Using PSI-blast, we obtained 501 sequences that align with the DHFR sequence from PDB 1RX7. With clustalW, we realigned these sequences

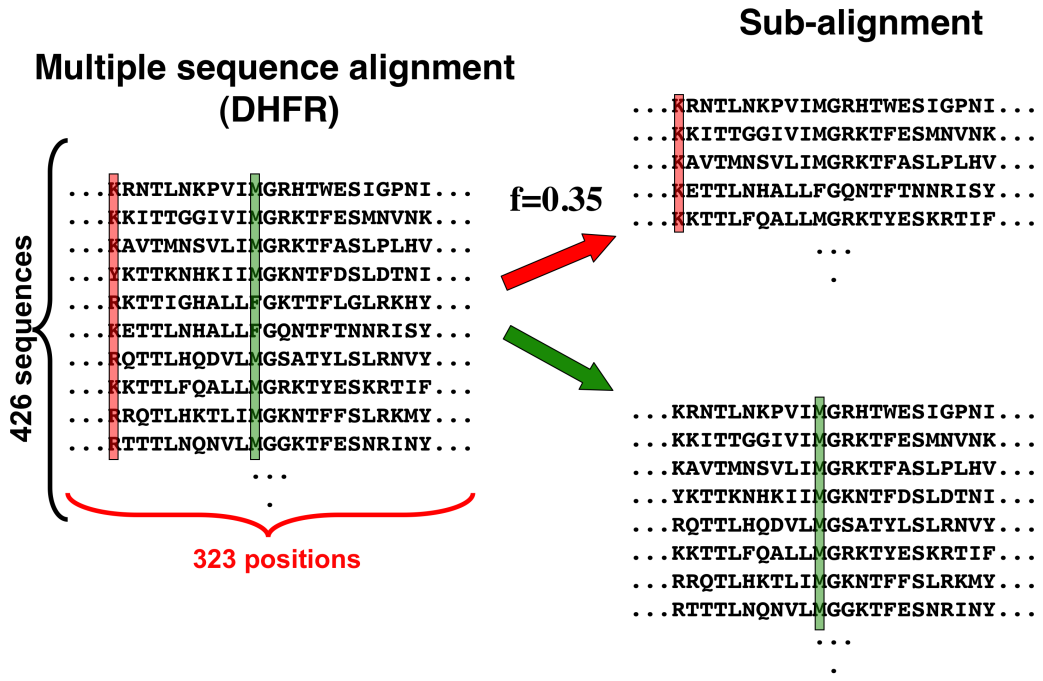


Figure 2.14: Multiple sequence alignment of DHFR family. 426 sequences are aligned, and the cutoff $f = 0.35$ is used to obtain sub-alignment.

and manually removed sequences that are too long or too short to cause long gaps in the MSA. The resulting 426 sequences were used to build up the multiple sequence alignment for SCA.

The subsets of the MSA were built up using the 426 sequences with a cutoff value of $f = 0.35$. According to central limit theorem, Eq. 1.3 and Eq. 1.4 must be satisfied when the size of the subalignment $N \cdot f$ is large enough. We took the difference of $\langle \overline{\Delta G} \rangle_f$ and σ_f^2, and found when $f < f_c = 0.35$, these differences change significantly, which indicates the large number law is no longer satisfied.

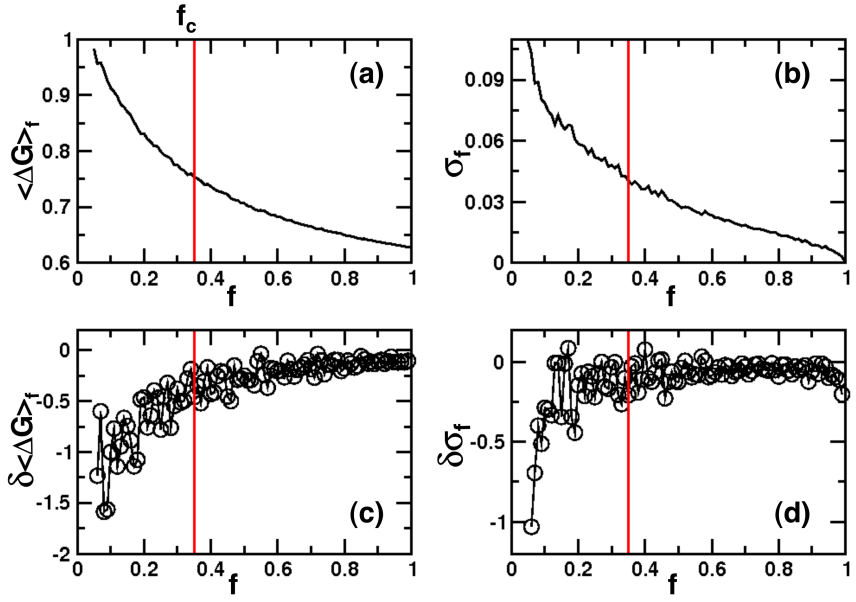


Figure 2.15: Central limit theorem is used to decide a cutoff value f for the subsets of MSA. $\langle\overline{\Delta G}\rangle_f$ and σ_f^2 are plotted in (a) and (b). In (c) and (d), the differences of $\langle\overline{\Delta G}\rangle_f$ and σ_f^2 are plotted. At $f < f_c = 0.35$, both $\langle\overline{\Delta G}\rangle_f$ and σ_f^2 start to deviate greatly from the values in MSA.

Chapter 3

Promoter melting triggered by bacterial RNA polymerase

3.1 Overall goals:

Transcription initiation is the first step in gene expression. The DNA-dependent RNA polymerase is the key enzyme in the transcription of the genetic information in all organisms [85]. It is essential to understand how RNA polymerase facilitates the transcription initiation. In recent years, structural studies have revealed the nature of the RNA polymerase initiation complexes [86, 87, 88, 89, 90, 40, 41, 91, 92]. To understand the dynamics of the transcription initiation, biochemical and generic experiments in various transcription conditions have been done to trapped RNA polymerase in the intermediate states [93, 37, 94]. Besides, single-molecule techniques, including single-molecule fluorescence (FRET) [95, 72], and magnetic tweezers (MT) [96] have been employed to probe the kinetic processes during the transcription initiation. The great challenge we face now is to connect the kinetics with the structural characterizations to understand the various conformational transitions in the initiation process. In this chapter, we focus on the transcription initiation triggered by bacterial RNA polymerase (RNAP), which shares considerable sequence and even more structural homology with its eukaryotic counterparts, e.g. RNA polymerase II (Pol II), and discuss recent progress in transcription initiation study achieved using both experimental and theoretical tools. Finally, I highlight some of the unsolved

questions and ongoing debates.

The outline of the chapter is the following: in section 3.2, we describe the transcription cycle of RNAP. In section 3.3, we focus on the transcription initiation step, and present the available structures of RNAP initiation complexes. In section 3.4, we review mutagenesis experimental results. In section 3.5, we present the results of a recent study of promoter melting using Brownian dynamics simulation. In section 3.6., we discuss the the fresh insights into the means by which RNAP forms transcription open complex (RPo) and conclude the simulation results. In section 3.7, we discuss various aspects of the Hamiltonian switching method.

3.2 Transcription cycle of RNAP

RNAP transcription cycle can be roughly divided into three major steps [97] (see figure 3.1), initiation, elongation and termination. All of the three steps are vastly complex, and include multiple intermediates. In the following, we describe the main intermediates in transcription cycle using RNAP as an example. During initiation, RNAP core-enzyme first associates with an co-enzyme called σ factor (Pol II requires multiple transcription factors) to form initiation-competent holoenzyme; σ factor is responsible for specific recognition of promoter DNA from a vast excess of non-promoter DNA by RNAP [98], and mediates the interaction of RNAP with promoter DNA and formation of a closed RNAP-promoter DNA complex ($R{\cdot}P_c$). $R{\cdot}P_c$ is in equilibrium with $R{\cdot}P_o$, a open RNAP-promoter DNA complex, in which the promoter DNA in vicinity of the transcription start site (from -10 to +2 site

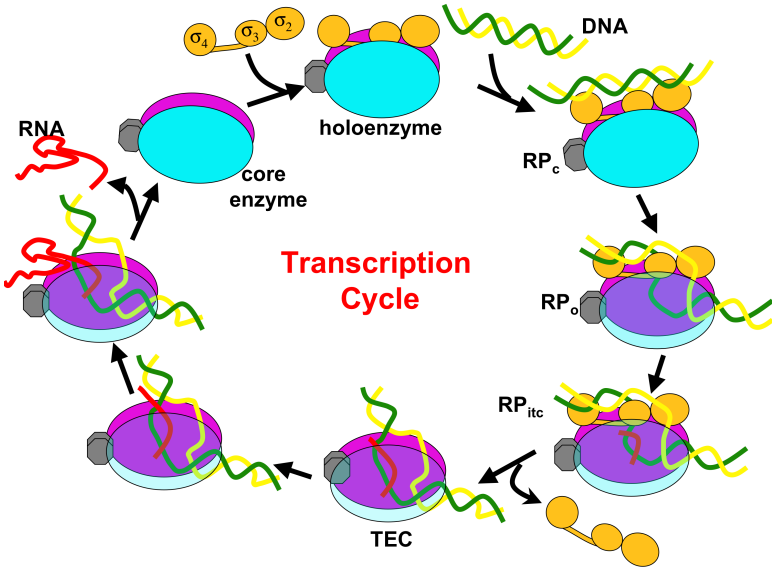


Figure 3.1: Transcription Cycle of RNAP

with +1 denotes the transcription start site) becomes single stranded and forms transcription bubble. $R{\cdot}P_o$ then enters abortive initiation in which multiple short transcripts (about 2-12 nucleotides in length) are synthesized and released. Abortive initiation continues until a transcript is elongated so transcription initiation can proceed. At this stage, σ factor is released and RNA transcription elongation complex (TEC) is formed. During elongation, nucleotides are added to TEC and the RNA transcript grows when RNAP translocates towards DNA downstream end with occasional backtrackings and pauses. At the end of elongation, the newly synthesized RNA is released and RNAP dissociates from the DNA and the transcription is terminated.

Among these highly regulated stages, the most complicated may be the initiation process because it involves promoter recognition, DNA unwinding and the formation of transcription bubble inside the RNAP, where RNA synthesis occurs.

The transcription initiation reaction pathway can be summarized as [99, 100]

$$\mathrm{R} + \mathrm{P} \rightleftharpoons \mathrm{R{\cdot}P_c} \rightleftharpoons \mathrm{R{\cdot}P_o} \rightleftharpoons \mathrm{itc_{\leq 12}} \rightarrow \mathrm{TEC}, \tag{3.1}$$

where R is RNAP, P is the promoter DNA (Fig. 3.5(a)), $\mathrm{R{\cdot}P_c}$ and $\mathrm{R{\cdot}P_o}$ are the closed and open complexes (Fig. 3.5(b)), respectively. $\mathrm{itc_{\leq 12}}$ is the abortive initiation complex with transcript size $\leq$ 12nt, and TEC is the transcription elongation complex.

3.3 Structures of RNAP

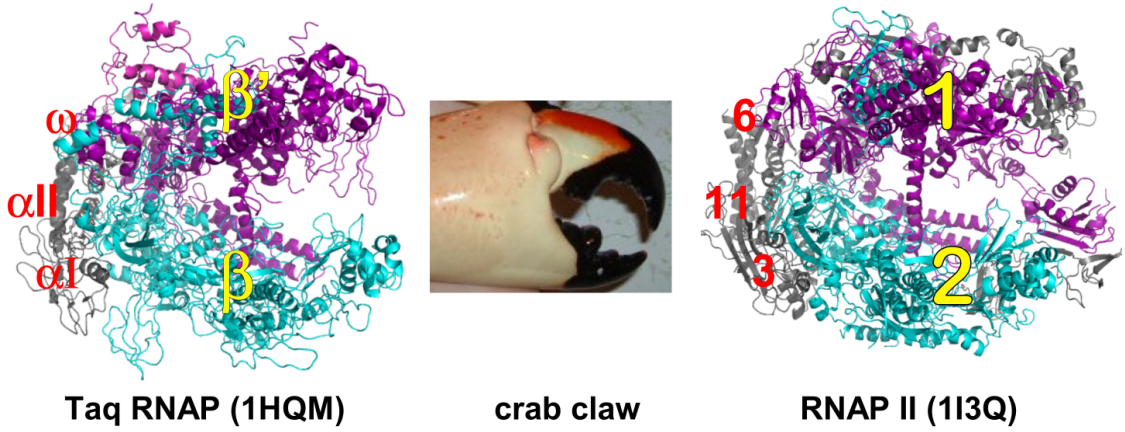


Figure 3.2: Top views of RNAP (left) and Pol II (right). Crystal structures for RNAP and Pol II are from Taq RNAP core enzyme(PDB 1HQM), and Pol II from yeast (PDB 1I3Q). The core enzymes of RNAP and Pol II have the crab-claw shape. In the middle, a crab claw is shown for comparison.

Transcription initiation triggered by RNAP is carried out by RNAP holoenzyme (holo), which comprises core RNAP (core) plus the initiation specific subunit σ [101, 98]. The core enzyme is a five-subunit complex $\alpha_2\beta\beta'\omega$, and has a crab-claw shape and a molecule weight of $\sim$ 400kDa [87]. Subunits β and β' are two "pin-

cers" of the crab-claw, whereas subunits αI, αII, and ω form the back wall of the crab-claw. From the back of to the tip of the claw, RNAP is about 150 $\mathring{A}$ long, 115 $\mathring{A}$ tall and 110 $\mathring{A}$ wide. Between the two pincers, an internal DNA binding channel of $\sim$ 27 $\mathring{A}$ in diameter exist, and the enzyme active site is located on the back wall of the channel. RNAP share the crab-claw shape structure with Pol II, which has a similar molecule weight of $\sim$ 500 kDa. In Pol II, the equivalent structural units for the pincers are subunits Rpb_2 and Rpb_1, and for the back wall are subunits Rpb_3, $Rpb_1 1$, and Rpb_6 (see Fig. 3.2). The active site of the two enzymes are highly conserved and the folds are essentially the same.

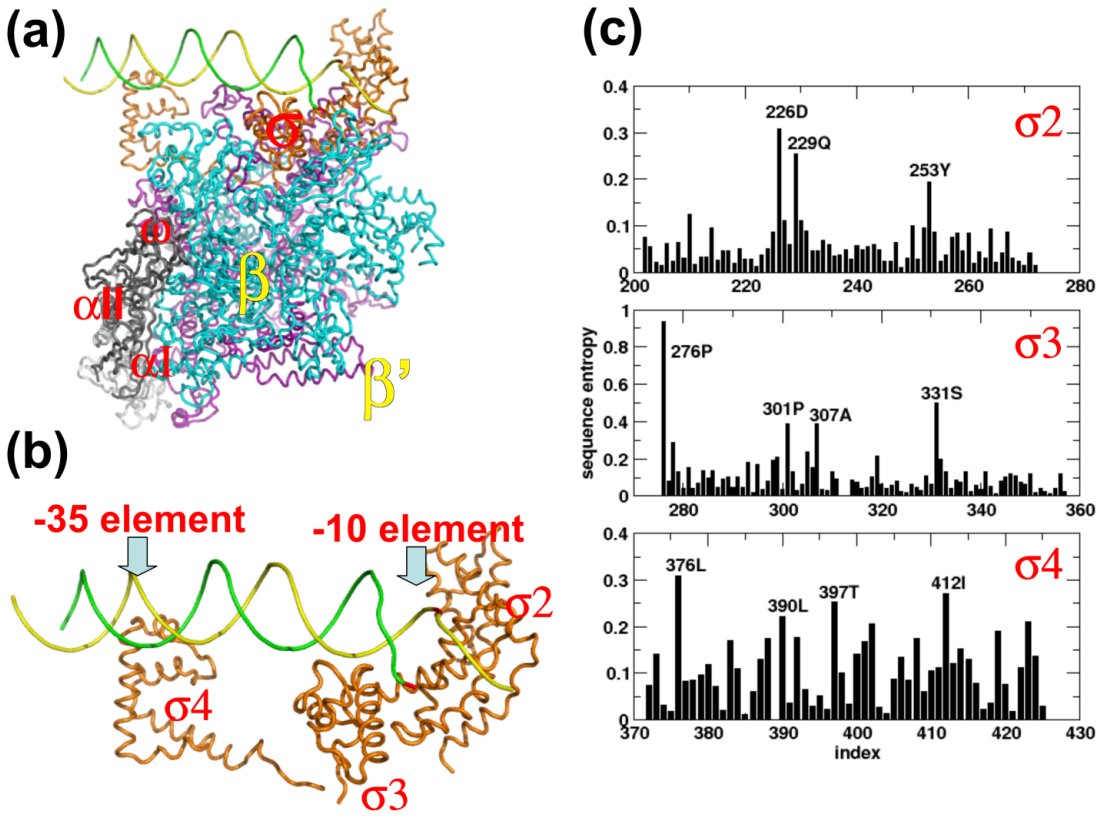


Figure 3.3: Coenzyme: σ factor

The σ subunit is the key regulator of bacterial transcription and is required for promoter-specific initiation of transcription [101, 98]. A primary σ^{70} factor (also called "housekeeping" σ factor), transcribing most of the genes in growing cell, comprises four major domains, $\sigma_{1.1}$, σ_2, σ_3, σ_4, connected by flexible linkers [98, 102]. In the holoenzyme, σ factor binds to the core and the complex is relatively stable through specific contacts between all of the four domains as well as the linkers to the core. On the other hand, the interaction between σ and the promoter DNA is formed between σ_2 and σ_4 and the -10 and -35 elements of promoter DNA (see Fig. 3.3 (b)-(c)). Two crystal structures, both from Darst group, give insight into the function of σ on binding of holo to promoter DNA [102, 91]. The first is a complex of isolated σ_4 with -35 element of promoter DNA, showing a 36^o bending of DNA around the -35 element upon binding to σ_4 [102]. The second is *T.Aquatics* RNAP holo complex with fork-junction promoter DNA, in which promoter DNA is double stranded from -41 to -12 and single stranded non-template DNA from -11 to -7 [91].

Genetic studies indicate that , σ_2 and σ_4 are the most conserved domains in σ [103]. $\sigma_{1.1}$ is poorly conserved in sequence but retains the acidic characteristics which helps to autoinhibit the promoter recognition by free σ factor. Highly conserved aromatic and basic residues in region 2.3 and 2.4 of σ_2 interact with -10 element, which suggest σ_2 is essential for σ function such as core binding, promoter DNA recognition and melting. σ_3 interacts with extended -10 element, and stabilizes the melted single-stranded DNA. It is concluded that $\sigma_{1.2}$-$\sigma_{3.1}$ contain all that is necessary and sufficient for basic σ factor functions.

Based on a variety of structures for RNAP, two groups modeled the holo in $R \cdot P_o$ [91, 40]. By fitting elements of B-form DNA (-60 to -37), DNA from the 2.4 $\mathring{A}$-resolution structure of the σ_4/-35 element complex, DNA from the RNAP-fork-junction DNA complex structure, and the ternary elongation complex model, Darst and coworkers determined the relative positioning of the promoter DNA inside the active site channel in $R \cdot P_o$ [91]. Ebright and coworkers used systematic FRET to determine the docking distance between promoter DNA and RNAP [40]. The structural models constructed by both groups are essentially identical except that Mekler *et al* were able to identify the position of $\sigma_{1.1}$ which was not present in available crystal structures of RNAP. In $R{\cdot}P_c$, $\sigma_{1.1}$ was found to locate in the active-site channel, but it was displaced outside in $R{\cdot}P_o$ [40].

3.4 Experimental results

RNAP structural models for $R{\cdot}P_c$ and $R{\cdot}P_o$ demonstrate major conformational changes of promoter DNA and RNAP. A number of experimental studies suggest that the promoter DNA is double stranded in $R{\cdot}P_c$ but melts between -12 and +2 in $R{\cdot}P_o$. For RNAP, because the diameter of RNAP active channel is narrower than the diameter of double-stranded DNA in both $R{\cdot}P_c$ and$R{\cdot}P_o$, entry of DNA into the RNAP must require conformational changes around the active-ste channel. On the other hand, $\sigma_{1.1}$ must move out from the active-site channel. The proposed melting of the promoter DNA raises a number of interesting questions that are difficult to answer using experiments alone. (1) What is the mechanism of RNAP induced DNA

melting? (2) What are the sequence of events in the conformational transitions from $R{\cdot}P_c \rightarrow R{\cdot}P_o$?

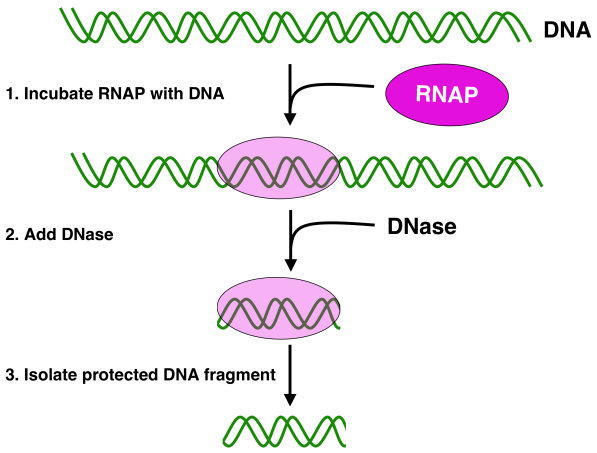


Figure 3.4: Adapted from Fig 13.6 in [104]

Using DNase and the single-strand selective reagent potassium permanganate ($KMnO_4$), foot-printing experiments (Fig. 3.4) are utilized to identify the intermediate species of RNAP in transcription initiation [93, 37, 94]. By varying transcription conditions such as temperature, concentration of Mg^{2+} or nucleotides, RNAP is found to establish partial opened transcription bubbles with strand separation occurs from -12 to sites upstream of +2. All of the partial opened transcription bubbles involve upstream region from -12, which suggested that promoter melting initiates from -10 region and propagates downstream [94].

Scanning force microscope images of the E. coli RNAP complex with lambda PL promoter reveal bending of DNA in $R{\cdot}P_o$ [42, 43, 44]. Along with the genetic study and footprinting experiments, the following picture of the transcription bubble formation has been postulated. Binding of the subunit $\sigma_{2.3}$ of the transcription

factor to the -10 element of the promoter DNA (Fig. 3.5(b)) results in local melting, which results in the nontemplate strand being stabilized by $\sigma_{2.3}$ (Fig. 3.5(b)). In addition, local melting renders the promoter DNA flexible, thus facilitating its entry into the active channel. The $R{\cdot}P_o$ structure further suggests that the promoter DNA bends into RNAP active channel to form the transcription bubble [42, 43, 44]. However, based on the crystal structure of RNAP holoenzyme from *Thermus thermophilus*, it was proposed that the β-subunit loop and σ-subunit α-helix obstruct the promoter DNA's pathway to the binding sites. It was argued that the β-subunit loop serves as a "gate" that allows only single stranded DNA to go through, and the "gate" stabilizes the DNA orientation to favor melting [92]. Although the structural models provide plausible hypothesis of the transcription bubble formation, dynamical studies are required to describe the conformational changes that accompany the $R{\cdot}P_c \rightarrow R{\cdot}P_o$ transition.

3.5 Theoretical approaches

We use the structural models for $R{\cdot}P_c$ and $R{\cdot}P_o$ [91] from bacterial *T. aquaticus* to address the following questions: (a) What are the steps in the transcription bubble formation, and (b) What is the nature of RNAP dynamics that enables the downstream dsDNA entry into the enzyme? To answer these questions, we performed Brownian dynamics simulations of the $R{\cdot}P_c \rightarrow R{\cdot}P_o$ transition using a coarse-grained self-organized polymer (SOP) model [35]. With the reduced description of the RNAP-DNA complex we performed detailed simulations of the kinetics

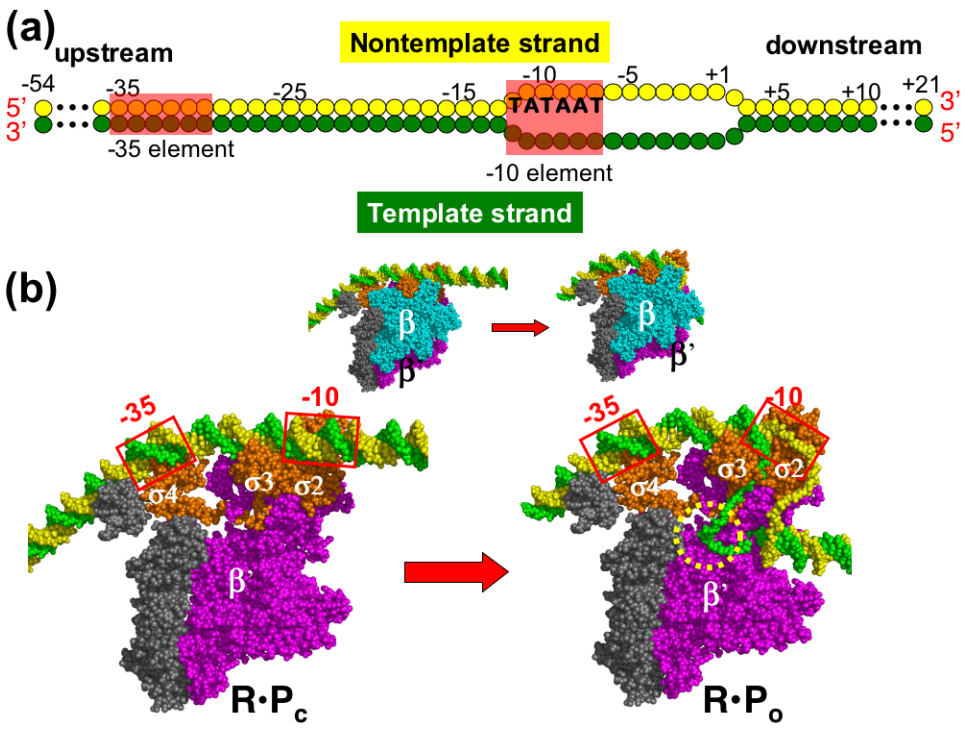


Figure 3.5: Structural models for the promoter DNA and RNAP. (a) Schematics of the base-pairing between the template (green) and non-template (yellow) strands of the promoter. Nucleotide positions are numbered relative to the transcription start site, +1. DNA segments that interact with RNAP, -35 and -10 elements, are in red shaded squares. (b) Structural models on top correspond to $R{\cdot}P_c$ (left) and $R{\cdot}P_o$ (right) and are color-coded as β cyan, β'-magenta, σ-orange, αI, αII, and ω-gray, DNA nontemplate strand-green, and template strand-yellow. The β subunit is removed to show the interactions of the promoter with the σ subunits (bottom left) and the transcription bubble structure on the bottom right, while on the top row it is shown in full opacity for reference.

of the $R{\cdot}P_c \rightarrow R{\cdot}P_o$ transition by generating a number of trajectories. Transcription bubble forms in three distinct steps. In step I, the region near the -10 element on the promoter DNA melts. In step II, the promoter DNA scrunches into the RNAP active channel to form a transcription bubble, while in step III the downstream DNA bends. Bending of the DNA is a result of downstream DNA relaxation, and occurs only after unwinding of the dsDNA. We also show that the widening of the channel needed to accommodate the dsDNA entry into the active channel requires

transient expansion of key structural elements in the β subunit of RNAP, which implies that the internal enzyme dynamics plays an important role in the $R{\cdot}P_c \rightarrow R{\cdot}P_o$ transition.

3.5.1 A network of contacts trigger the promoter melting.

The structures of the RNAP-DNA complexes used in this work [91] have 3122 residues and 150 nucleotides. The holoenzyme has 6 subunits: αI (Ala6-Glu229), αII (Ala6-Phe225), β (Ala2-Ala1116), β' (Ala3-Ala1499), ω (Ala2-Ala93), and σ (Ala93-Ala438) (Fig. 3.5). The promoter DNA has 75 base pairs, -54 to +21, labeled with respect to the transcription start site +1 (Fig. 3.5a). In $R{\cdot}P_c$, the promoter DNA "sits" on the top of RNAP (Fig. 3.5b) and forms stable interactions between the σ subunit and the -10 and -35 regions. Despite the low-resolution nature of the models, the dynamics reveal key structural changes that occur during the $R{\cdot}P_c \rightarrow R{\cdot}P_o$ transition. Several contacts between DNA -10 element and the subunit σ of RNAP rupture, in particular, the contacts involving nucleotides -12 to -8 on the template strand (Fig. 3.6). Formation of contacts between nucleotides -3 to +5 on the non-template strand and the β subunit (Fig. 3.6), nucleotides -9 to +1 on the template strand and the β subunit (Fig. 3.6), and nucleotides +1 to 7 on the template strand and the β' subunit(Fig. 3.6) stabilize the $R{\cdot}P_o$ state.

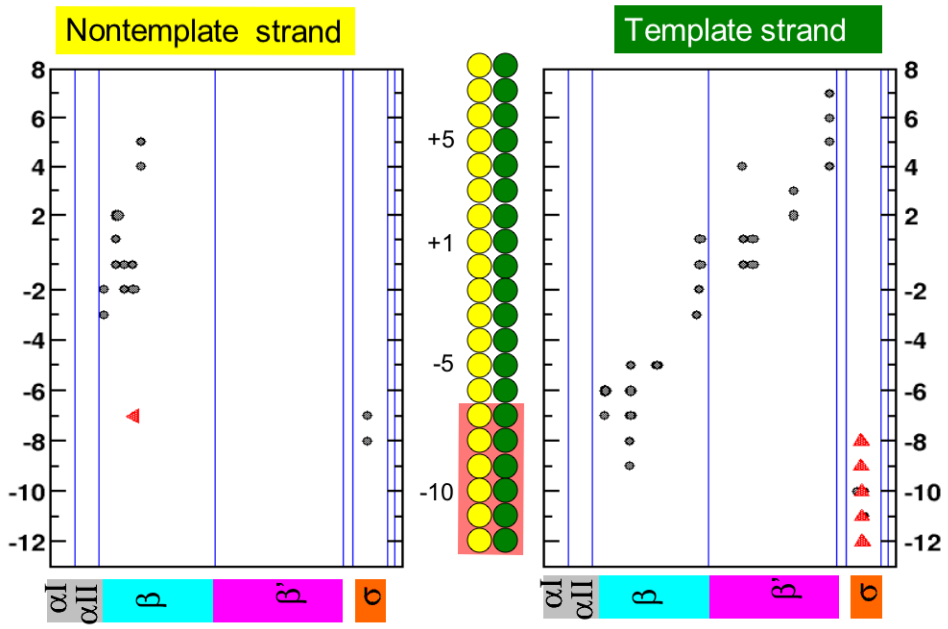


Figure 3.6: Contact map of RNAP complex with DNA. The vertical axis gives the nucleotide index in the promoter DNA and the horizontal axis is the domain partitioning of RNAP. The 5 subunits of RNAP, αI, αII, β, β', and σ, are labeled underneath the residue axis. The contacts between RNAP and the DNA nontemplate strand are shown on the left panel and the contacts between RNAP and the DNA template strand are shown on the right panel. In both cases, red triangles represent the contacts that exist in R·P_c but not in R·P_o, and the gray circles represent the contacts that exist in R·P_o but not in R·P_c.

3.5.2 $R{\cdot}P_c \rightarrow R{\cdot}P_o$ transition trajectories partition into fast (efficient) and slow (inefficient) tracks:

The global nature of the R·P_c→R·P_o transition is monitored by the time dependent changes of the root mean square deviation Δ_C(t) (Δ_O(t)) of the DNA with respect to the closed (open) value. The transition times, τ_ms, calculated using $|\Delta_C(\tau_m) - \Delta_O(\tau_m)| < \epsilon = 0.1$ $\mathring{A}$, were used to partition the set of 30 trajectories into fast and slow processes (Figs. 3.7(a-c)). In the fast routes, Δ_C(t) (Δ_O(t))

increases (decreases) very rapidly (Fig. 3.7 (a)), which indicates (see below) that upon opening of the DNA base pairs the transcription bubble forms efficiently. In contrast, in the slow routes, long-lived metastable states, which are indicated by a plateau in Δ_{C}(t) (Δ_{O}(t)) (Fig. 3.7 (b)), are populated. The base pairs open by a complicated pathway resulting in a decreased efficiency in the transcription bubble formation.

What is the origin of the differences between the fast and slow trajectories? To answer this question, we examined the time-dependent conformational changes of the promoter DNA, which we describe using the distances, $d_i(t)$, between the two complementary nucleotides of each base pair in the promoter DNA. Here, $d_i(t) = |\vec{r}_i^{T}(t) - \vec{r}_i^{NT}(t)|$ where $\vec{r}_i^{T}(t)$ and $\vec{r}_i^{NT}(t)$ are the positions of the i^{th} nucleotide on the template and the non-template strands, respectively. Since the transcription initiation site is at $i = +1$ and the transcription bubble forms between $i = -12$ and $+2$ [99] , we computed $d_i(t)$ for four representative base pairs, -11, -7, -3, +2 to describe the transcription bubble formation. In the R·P$_{\mathrm{c}}$ state, $d_i(t) \approx 11$ $\mathring{A}$ for all of the four base pairs, but in R·P$_{\mathrm{o}}$ state, $d_i(t) \approx 18$ $\mathring{A}$, 48 $\mathring{A}$, 53 $\mathring{A}$, and 28 $\mathring{A}$ for -11, -7, -3, and +2 base pairs, respectively.

3.5.3 Promoter DNA unzips sequentially from -10 element in the fast track.

Analysis of $d_i(t)$ in all the fast trajectories shows a consensus sequence of events during the R·P$_{\mathrm{c}}$→R·P$_{\mathrm{o}}$ transition. Base pair -11 opens first at t∼10 μs, which is

followed by disruption of interactions in -7 at $t\sim16\mu$s. In both cases the equilibrium values corresponding the structure in R·P$_o$ is reached rapidly (Fig. 3.8). At $t\sim24\mu$s, the -3 pair rips and the distance between the nucleotides attains the value in the R·P$_o$ state. The distance $d_{+2}(t)$ fluctuates between 11 $\mathring{A}$ and 30 $\mathring{A}$, and reaches 30 $\mathring{A}$ immediately after base pair -3 opens. The opening of the upstream base pairs favor the rupture of the downstream neighbors, which establishes that the base pairs from -12 to +2 rupture abruptly in a sequential manner by an unzipping mechanism. Complete analysis of all the $d_i(t)$s shows that sequential unzipping starting from the -10 site leads to rapid transcription bubble formation.

3.5.4 Initial base pair opening away from the -10 element results in slow $\mathbf{R{\cdot}P_c{\rightarrow}R{\cdot}P_o}$ transition.

Although promoter recognition sequences are localized in the -35 and -10 regions, melting of the base pair can in principle occur stochastically. However, as seen in the fast trajectories for the transcription bubble to form efficiently, rupture must start from the -10 element. In some trajectories, multiple base pairs melt nearly simultaneously in a non-sequential process, which greatly impedes the transcription bubble formation. Fig. 3.8 shows that in one of the slow trajectories, rupture of base pairs -3 and +2 compete with the opening of the base pair -7. At $t\sim$ 40μs, base pair -7 remains intact while the downstream base pairs -3 and +2 rip as seen in the increase of $d_{-3}(40\ \mu\text{s})$ and $d_{+2}(40\ \mu\text{s})$ from 11 $\mathring{A}$ to 40 $\mathring{A}$ and 11 $\mathring{A}$ to 50 $\mathring{A}$, respectively (Fig. 3.8). At longer times, the base pairs -3 and +2 backtrack as

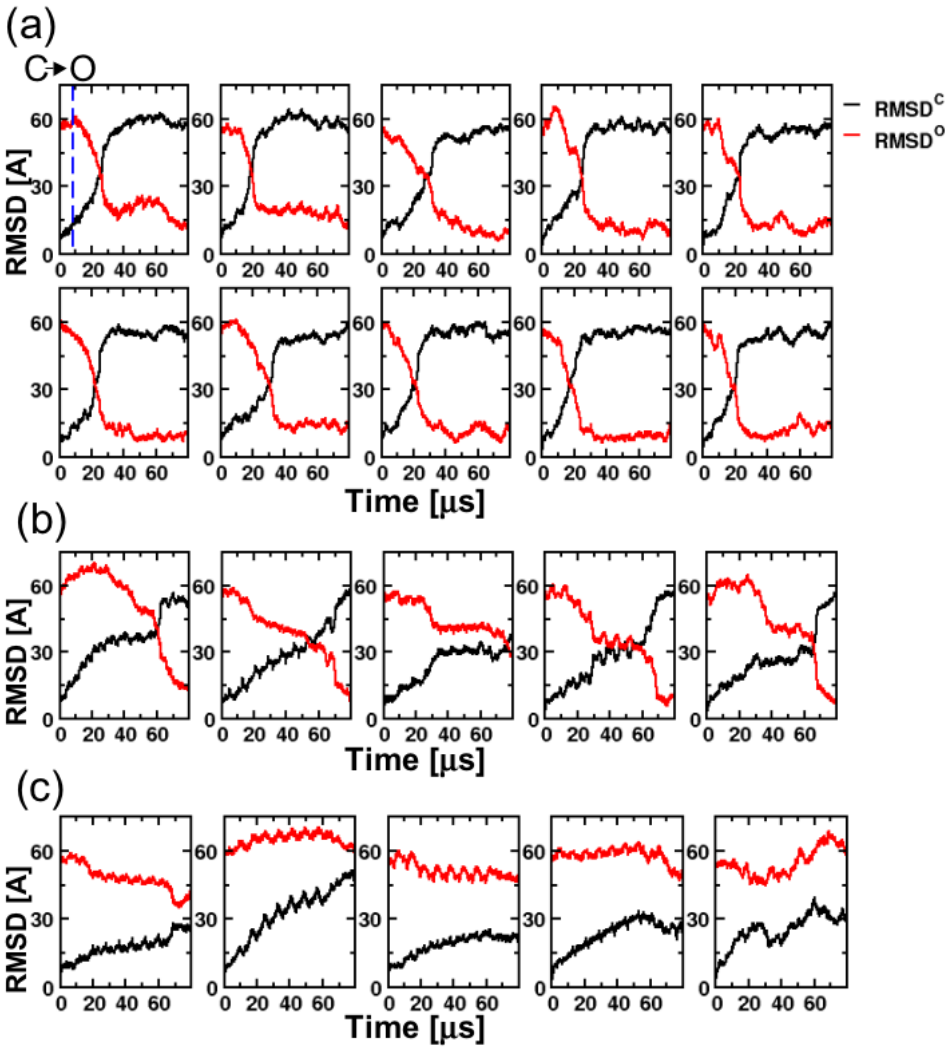


Figure 3.7: Diversity of routes during the R·P_c→R·P_o transition assessed by root mean square derivation as a function of time. (a) Ten fast trajectories with mean relaxation time $\tau_m < 30\mu$s (see text for defination of τ_m). (b) Five slow trajectories with $\tau_m > 50\mu$s. (c) Five super slow trajectories with $\tau_m >> 50\mu$s. In all panels black (red) curve gives RMSD with respect to R·Pc (R·Po).

shown by the decrease in d_{-3} and d_{+2} to 11 $\AA$, the value in the R·P_c state, resulting in the "reseting" of the initial state, which leads to melting of the -7 base pair. Subsequently, melting of the base pairs occur sequentially, which is manifested in the increase of $d_i(t)$ to the values in the R·P_o state.

We also observed trajectories in which DNA retains double stranded form for

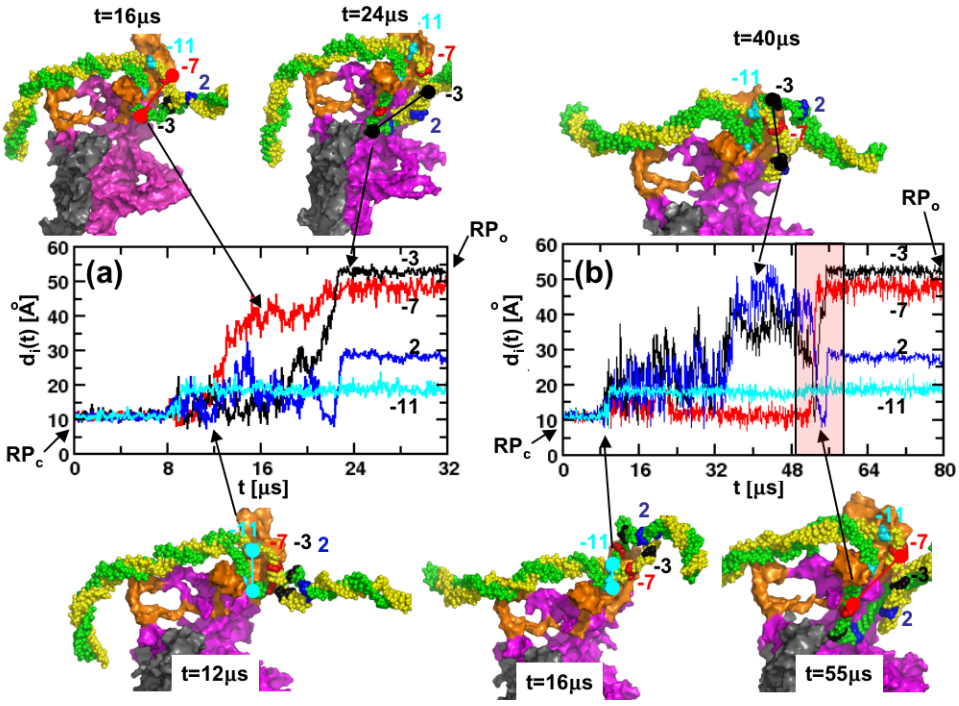


Figure 3.8: Dynamics of transcription bubble formation. (a) Distance between complementary nucleotides of the base pairs in the bubble region as functions of time for a representative fast trajectory. Cyan, red, black and blue lines correspond to $d_i(t)$ for base pairs -11, -7, -3 and +2, respectively. Structures of the RNAP complex at t=12μs, 16μs, and 24μ during the development of the transition bubble are shown. (b) Same as (a) except the results are for one of the slow trajectories. The reforming of prematurely opened base pairs -3 and 2 (black and blue curves respectively) is highlighted in the shaded region. The structures sampled during the R·P$_c$→R·P$_o$ transition are hightlighted. In both (a) and (b), the arrows indicate the starting and ending states.

a relatively long time (3.9). Analysis of the distances d_it show that all of the four base pairs -11, -7, -3 and +2 melts all at once, however, none of the base pairs is able to further melt and separate. The all or none behavior of the bubble formation is inefficient in forming transcription bubble.

Taken together, the data show that the rapid formation of the transcription bubble requires sequential unzipping from the -10 element. The importance of the -10 element in initiating the base pair opening can be related to its interaction with

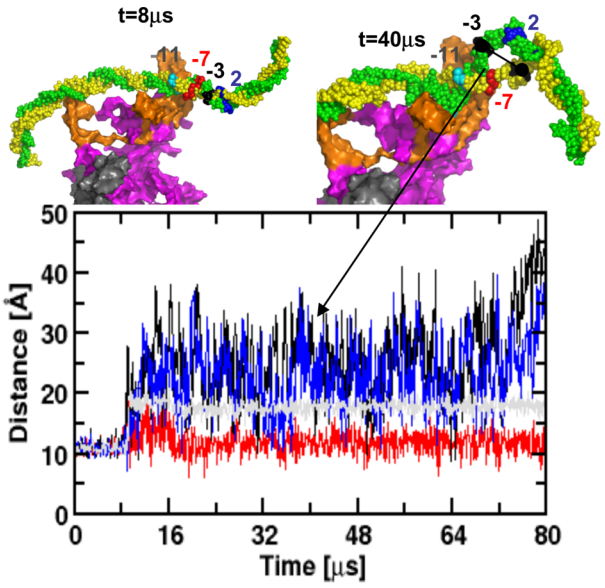


Figure 3.9: Distance between complementary bases in the bubble region as functions of time for a representative super slow trajectory. Snapshots of RNAP complex structure at t=8μs and 40μs are shown. The color scheme is the same as the in Fig.3.8.

the RNAP through the subunit σ, which is well known to be essential for DNA melting and subsequent stabilization of the initial transcription bubble [97].

3.5.5 The transcription bubble formation in fast trajectories occurs in three steps.

We used the time-dependent distance changes between the i^{th} (j^{th}) nucleotide on the promoter sequence and j^{th} residue on RNAP, $d_{ij}(t) = |\vec{r_i}(t) - \vec{r_j}(t)|$, $\vec{r_i}$ ($\vec{r_j}$) is the position of the i^{th} nucleotide (residue), to identify three steps in the transcription bubble formation process. The overall dynamics associated with the

bubble formation is assessed using $d_{-1}(t)$, which on an average (black line in Fig. 3.10 (a)) occurs in about (30-40) μs. However, there is heterogeneity in the individual trajectories. Dissection of the events leading to the increase in $d_{-1}(t)$ from about 11 $\AA$ at t=0 to 55 $\AA$ at t$\approx$ 35μs shows that transcription bubble formation occurs in three major steps.

Step I: The -10 element melts locally. The decreases in $d_{-12',Mg^{2+}}(t)$ (the prime refers to the base pair number on the NT strand) (Fig. 3.10 (b)) shows that the nontemplate strand moves towards the active region of the enzyme (amino acids around the Mg^{2+} ion). In this stage, $d_{-12,+21}$ also decreases (Fig. 3.10 (b)), but in contrast to $d_{-12',Mg^{2+}}(t)$, the changes in the conformations of +21 base pair relative to the RNAP continues to evolve throughout the transcription bubble formation process (see below). Compared to the time needed for the bubble opening (Fig. 3.10 (a)), the characteristic time associated with the decrease of $d_{-12',Mg^{2+}}(t)$ (t$\sim$16 μs) is short, and is the major feature of step I. Although the template strand of the promoter DNA also moves synchronously in this step (data not shown), the dynamics of such a process lasts over the entire duration of bubble formation(at t$\sim$35 μs). Therefore, we do not consider the movement of the template strand as a major feature of step I. The correlated movement of the nontemplate and the template strands is consistent with the assumption that local melting of DNA in the -10 element renders it flexible. In addition, the observed stabilization of the nontemplate strand in step I agrees with the experimental finding that the conserved aromatic residues of subunit σ are positioned to recognize and stabilize the exposed nontemplate strand [105, 106, 107]. Loss of base pair interaction upon melting of

the -10 base pairs (TATAAT) results in the formation of favorable interaction of adenine or thymine nucleotide with the aromatic side chains (Phe248, Tyr253, and Trp256) on σ2.3 subunit as well as electrostatic interactions between DNA and the enzyme.

Step II: DNA scrunches into the RNAP active-site channel and forms a bubble. The transcription bubble starts to grow from the -10 to +2 base pair as the promoter sequence unzips (Fig. 3.10). The bubble region quantified in terms of center, $d_{-1}(t)$ increases from 11 $\mathring{A}$ to 55 $\mathring{A}$, the value in the R·P_o state, which implies that in this step unwinding of the promoter DNA and strand separation occurs. Meanwhile, the distance between the -10 element and the downstream edge of the promoter DNA (base pair +21) decreases from 110 $\mathring{A}$ to 55 $\mathring{A}$ (see the V-shape $d_{-12,+21}$ in fig. 3.10(c)). The observed decrease in Fig. 3.10 (c) during step II is reminiscent of the scrunching mechanism (SM) proposed for the R·P_o→R·P_{itc} based on single molecule FRET experiments. According to SM, in each cycle of the abortive initiation, RNAP pulls the downstream DNA into the active channel and past its active center without substantial change in the enzyme conformation [96, 95]. Although the R·P_{itc} transition formation occurs only after the completion of the R·P_c→R·P_o transition, the scrunching mechanism appears to also drive the formation of the transcription bubble (see below).

Step III: Downstream DNA bends. In the final stage, the downstream DNA bends to release the strain accumulated during the promoter DNA opening. The strain release results in the increase in $d_{-12,+21}(t)$ after the decrease in step II (Fig. 3.10 (c)). In addition, DNA bends and kinks in the downstream region,

which is captured by the time-dependent changes in the angle, $\Psi(t)$ (Fig. 3.10 (d)). The angle $\Psi(t)$ formed between the interacting sites localized at -35, -12, and +21. which is relatively constant during step I and II, decreases in step III from $\sim 150^{o}$ to $\sim 100^{o}$, which is the value in the RP_o state [42, 43, 44]. The substantial change in $\Psi(t)$ results in DNA bending.

3.5.6 Scrunching is a universal mechanism in initiation:

To illustrate whether the promoter DNA scrunches into RNAP in the $R{\cdot}P_c \rightarrow R{\cdot}P_o$ transition, we calculated the time-dependent distance changes, $d_{ij}(t)$, for the same sites that are labeled in the FRET study [95, 72, 108, 109, 110] (Fig. 3.11). The conclusion that during the $R{\cdot}P_c \rightarrow R{\cdot}P_o$, the promoter DNA is scrunched follows the arguments used in the experimental study [95]. From the decrease in the distance between the leading edge of RNAP (residue 366 of σ) and the +20 site in downstream DNA (Fig. 3.11a), it follows that RNAP translocates relative to downstream DNA. The lack of distance changes $d_{\sigma 569,-39}(t)$ and $d_{\sigma 569,-20}(t)$ between the upstream edge of RNAP (σ569) and upstream sites in DNA (-39 and -20) during the course of $R{\cdot}P_c \rightarrow R{\cdot}P_o$ transition (Figs. 3.11 (b) and (d)) implies absence of translocation of RNAP relative to upstream promoter sites. Similarly, $d_{\sigma 366,-20}$ is unchanged during the $R{\cdot}P_c$ to $R{\cdot}P_o$ transition, which implies that the leading edge of RNAP is stationary with respect to upstream DNA. Direct measurements of distance changes between site -15 and +15 on the promoter DNA shows that $d_{-15,+15}(t)$ decreases by about 5 $\mathring{A}$ when the $R{\cdot}P_c \rightarrow R{\cdot}P_o$ transition is completed.

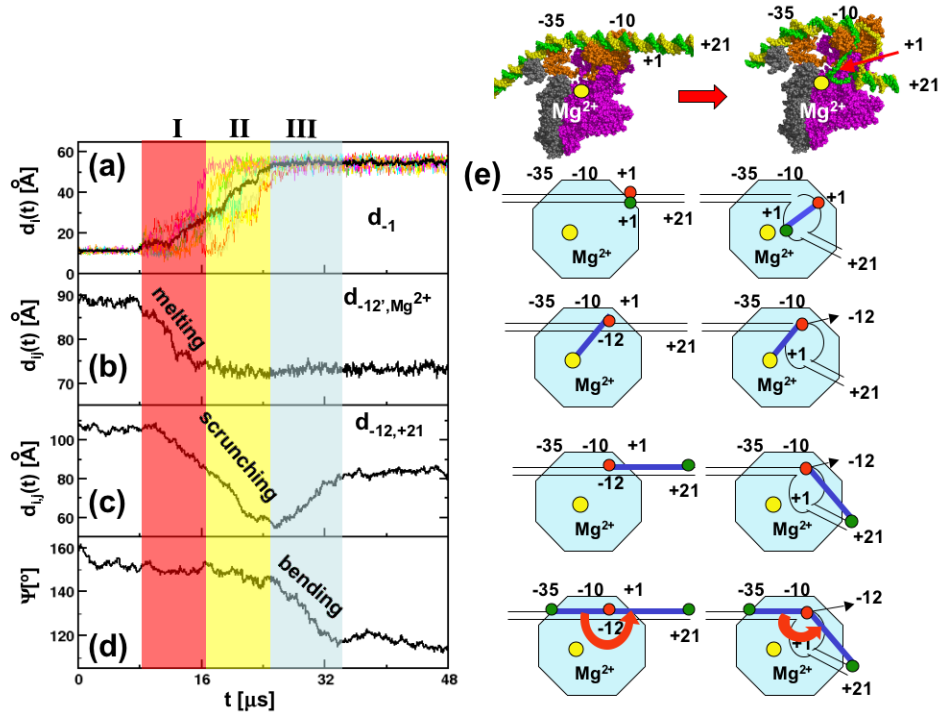


Figure 3.10: Three steps (in shaded colors) in the transcription bubble formation in the fast traectories. (a) The time-dependent increase in $d_{-1}(t)$ is used to assess the growth of the transcription bubble. The black line shows $d_{-1}(t)$ averaged over several trajectories, and the individual traces represent the data for a few trajectories. (b) Average changes in the distance between the nucleotide -12 on the non-template strand and Mg^{2+} as a function of time. (c) Ensemble average changes in $d_{ij}(t)$ for i=-12 and j=+21 as a function of t. (d) The time-dependent change in the angle Ψ between the interaction centers at -35, -10, and +21 shows that only after transcription bubble forms (see (a)), the dsDNA bend (Ψ decreases in step III, which is shaded in light blue color). (e) Schematic representation of dynamical processes in (a)-(d). The sketches from top to bottom correspond to bubble growth ($d_{-1}(t)$ increase), promoter melting (decrease in $d_{-12,Mg^2+}(t)$), DNA scrunching (decrease in $d_{-12,+12}(t)$), and dsDNA bending (decrease in $\Psi(t)$). On top right are shown the front view of R·P_c (left) and R·P_o (right) structures without the β subunit and a few relevant nucleotides are labeled for reference.

These results show that, to a large extent, RNAP is fixed on the promoter while the distances between downstream edge of DNA decrease. The "V" shape portion of $d_{\sigma 366,+20}(t)$ and $d_{-15,+15}(t)$ (Fig. 3.11 (a) and (e)) shows that the distance changes associated with leading edge and the specific sites on DNA do not change mono-

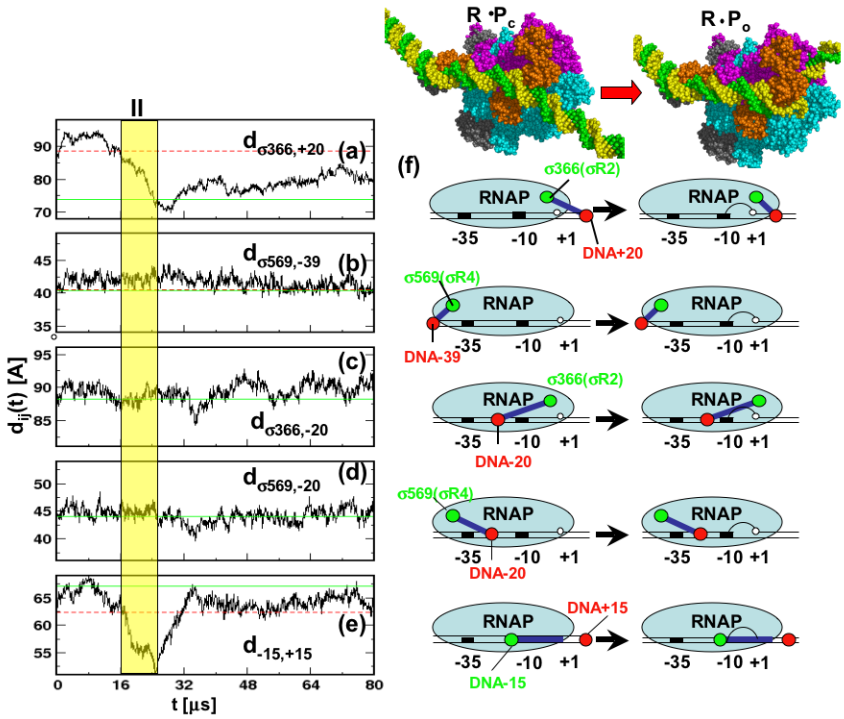


Figure 3.11: Dissection of DNA scrunching during Step II. Structures on the top right, top view of R·P$_c$ (left) and R·P$_o$ (right) structures are shown for reference. The distances between nucleotides of the promoter DNA and RNAP as function of time, $d_{ij}(t)$ are on the left. In (a) - (d) the red dashed lines (solid green lines) are d_{ij} values in R · P$_c$ (R · P$_o$) state. (a) Ensemble averaged distance between nucleotide +20 and the leading edge residue σ366 (numbering as in *E. coli).* (b) Change in the distance between the trailing head residue σ569 and nucleotide -39 as a function of t. (c) Time-dependent change in the distance between the leading edge residue σ366 and the upstream nucleotide -20. (d) Distance between the trailing edge residue σ569 and nucleotide -20 as a function of t. (e) Average changes in $d_{-15,+15}(t)$ as a function of t. (f) The sketches from top to bottom capture the distance changes shown on the left. Structures on the top right, top view of R·P$_c$ (left) and R·P$_o$ (right) structures are shown for reference.

tonically in R·P$_c$→R·P$_o$ transition. The distance changes between labeled sites on promoter DNA and changes in $\Psi(t)$ (Fig. 3.10 (c) and (e)) show that the "V" shape of the distances is largely due to the contraction (or scrunching) of the DNA in downstream region. The subsequent increase in the variables shown in Figs. 3.11 (a) and (e) is due to the DNA bending.

3.5.7 Active channel of RNAP opens by a "rope-swing" mechanism.

The dynamics of $R{\cdot}P_c \rightarrow R{\cdot}P_o$ transition shows how the holoenzyme melts the promoter sequence, and leads to efficient transcription bubble formation. Despite the enhanced flexibility of the melted region of DNA, it remains unclear how the promoter gains access to the binding sites in RNAP. The main DNA binding channel in the RNAP holoenzyme (10-12 $\mathring{A}$ in diameter for *T. thermophilus* and 16-19 $\mathring{A}$ for *T. aquaticus*) is narrow compared to the 22 $\mathring{A}$ diameter of the dsDNA. [41, 92]. Based on static structure alone, two models have been proposed to explain how the RNAP channel expands to accommodate the dsDNA. One possibility is that the repulsive interactions between the negatively charged σ1.1 domain and the dsDNA result in the opening of the channel [97]. Alternatively, using high resolution structure of the holoenzyme in the closed state, it has been suggested that the protruding region of the β-subunit could fit into dsDNA major groove [92]. However, in absence of the crystal structure of the intermediate states during the transcription bubble formation process or evidence from time-resolved experiments, the mechanism of opening of the main channel remains elusive. Our simulations of the transcription bubble formation process suggest that channel opening occurs by a "rope-swing" mechanism (see below) is triggered by transient movements of the structural elements of the β subunit.

Analysis of the structures reveals that Ala132-Ser387 region of the β-subunit makes extensive contacts with the promoter DNA in $R{\cdot}P_o$ but not in $R{\cdot}P_c$, which suggests that this region (referred to as the Rope-Swing (RS) domain) must play an

essential role in the channel widening process. To dissect the kinetic role of the RS domain we partition it into the N-terminal globular Swing (S) domain (Ala132-Arg334) and the Rope (R) domain corresponding to the C-terminal helix-turn-helix motif (Thr335-Ser387) (Fig. 3.12 (a)). The R domain is the analogue of the helix-loop-helix fragment in *T. thermophilus* RNAP [92]. The widening of the DNA binding channel is facilitated by the displacement of the RS domain as shown by time-dependent changes in a number of quantities associated with the R, the S domain and the β' subunit. The closest distance, $R(t)$, between the R and S domains, whose value is 14 $\mathring{A}$ in R·P_c, increases past the value in the final R·P_o state over a "gating" time interval $\Delta\tau_w$ that ranges from 19μs to about 28μs (Fig. 3.13 (b)). Over the gating time interval $\Delta\tau_w$ the changes in the distance $d_{AB}(t)$ (A is the center of mass of the RS domain and B is the center of mass of the residues Ala105-Ala499 of the β' subunit) as a function of t shows (Fig. 3.12 (c)) a total increase of 7 $\mathring{A}$ during the R·P_c→R·P_o transition. The 7 $\mathring{A}$ increase, which occurs transiently (Fig: 3.12 (b) and (c)), is sufficient to enable entry of the dsDNA. The change in $d_{AB}(t)$ is also accompanied by an increase of about 8^o in the angle, $\Omega(t)$ (Fig: 3.12 (d)), between A, O, B (O is the center of mass of residues Glu1264 - Lys1426 in the β' subunit) at $t \approx$24μs, which coincides with the increase in $d_{AB}(t)$ (Fig. 3.12 (c)). Both $d_{AB}(t)$ and $\Omega(t)$ decrease after DNA enters RNAP, which results in a tight grip on the dsDNA. Surprisingly, the distance $d_{OB}(t)$ is almost constant (Fig. 3.12 (c)). The totality of the results show that channel expansion occurs predominantly by swinging of the RS domain with the β' subunit being stationary.

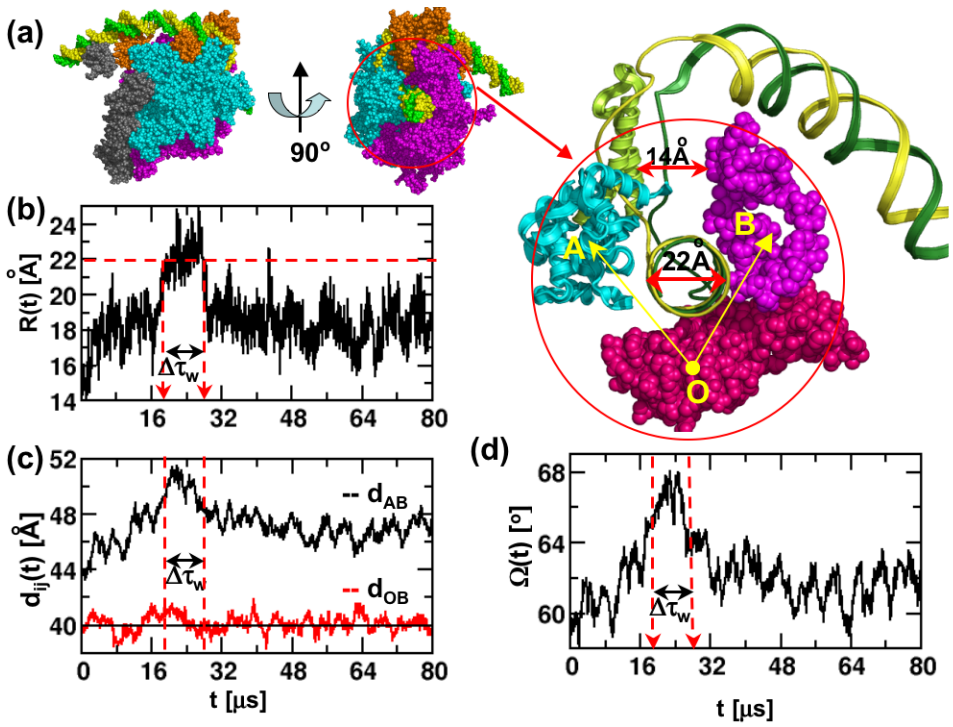


Figure 3.12: Dissection of DNA scrunching during Step II. Structures on the top right, top view of R·P_c (left) and R·P_o (right) structures are shown for reference. The distances between nucleotides of the promoter DNA and RNAP as function of time, $d_{ij}(t)$ are on the left. In (a) - (d) the red dashed lines (solid green lines) are d_{ij} values in $R \cdot P_c$ ($R \cdot P_o$) state. (a) Ensemble averaged distance between nucleotide +20 and the leading edge residue σ366 (numbering as in *E. coli).* (b) Change in the distance between the trailing head residue σ569 and nucleotide -39 as a function of t. (c) Time-dependent change in the distance between the leading edge residue σ366 and the upstream nucleotide -20. (d) Distance between the trailing edge residue σ569 and nucleotide -20 as a function of t. (e) Average changes in $d_{-15,+15}(t)$ as a function of t. (f) The sketches from top to bottom capture the distance changes shown on the left. Structures on the top right, top view of R·P_c (left) and R·P_o (right) structures are shown for reference.

To obtain additional insights into the channel widening mechanism, we calculated the dynamics of the molecular contacts between residues in the RS domain and the nucleotides in the DNA. Analysis of the structural changes that occur during the approximate time windows ($\Delta\tau_I$ ($8\mu s$ to $17\mu s$); $\Delta\tau_{II}$ ($17\mu s$ to $24\mu s$); $\Delta\tau_{III}$ ($24\mu s$ to $34\mu s$) characterizing the three steps and during $\Delta\tau_w$ provides a molecular picture

of the fate of the DNA during the transcription bubble formation. The interactions between upstream nucleotides in the non-template strands and the residues in the σ and β subunits are fully developed in τ_I (Figs. 6a1 and a2). For example, contacts between -8′ and 252H (chain H is in the σ subunit) and -2′ and 349C (chain C is in the β subunit) are fully developed in $\Delta\tau_I$. Similarly, certain contacts involving the nucleotides in the template strand and the enzyme are also fully formed in $\Delta\tau_I$.

In the first half of Step II, prior to channel opening, the promoter DNA continues to melt from -6 to -3 and the single stranded DNA scrunches into RNAP to contact residues on β subunit (Figs. 6(a2) and (b2)). Simultaneously, the template strand (-2 to +2) develops interactions with residues in the S domain that are absent (non-native) in the R·P$_o$ state (Figs. 6(c) and (d)). The formation of these transient non-native interactions leads to build up of strain in the transcription bubble (see Fig. 6e), which enables the channel to open (see the rise in $R(t)$ shown in Fig. 5b and simultaneous increase in the strain energy in Fig. 6e). In the second half of the scrunching process, the promoter DNA melts from -2 to +2 and interactions with the β and β' subunits fully develop. At the end of Step II ($t \approx 24\mu s$), which is roughly in the middle of $\Delta\tau_w$, the entire transcription bubble (-10 to +2) is formed and all the distances associated with the DNA have reached the values in the R·P$_o$ state.

In the first stages of Step III, the downstream dsDNA (+3 to +21) bends (see Fig. 3d showing the maximum decrease in $\Psi(t)$ occurs in Step III). Bending of the downstream DNA enables entry into the channel, and also results in the decrease of accumulated strain (see the decrease in the repulsive energy in Fig. 6e).

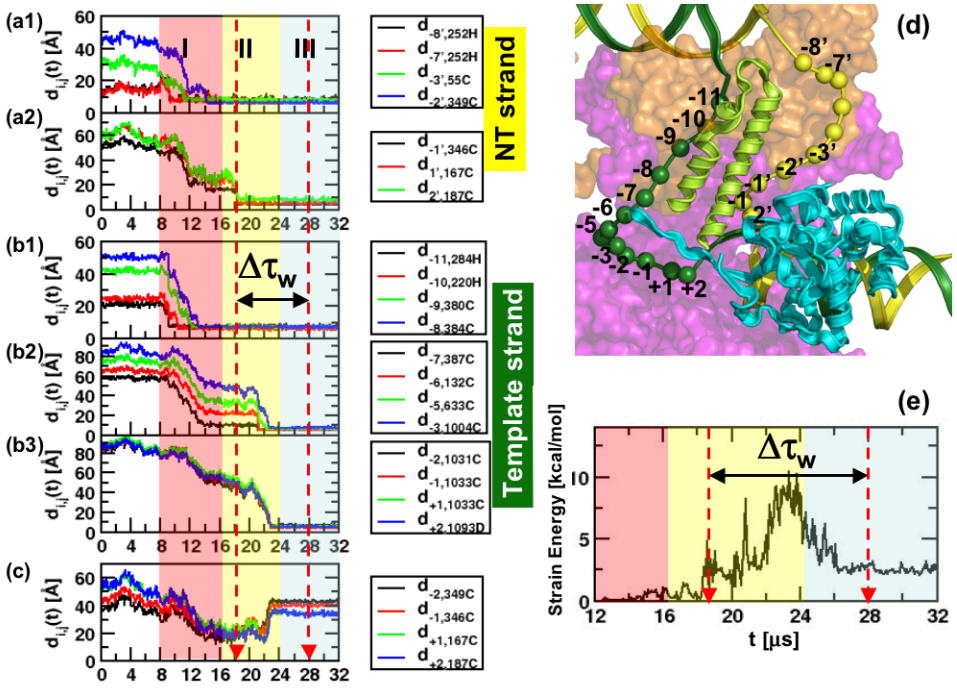


Figure 3.13: Molecular details of channel opening mechanism. (a1) Evolution of the distances in a fast trajectory between nucleotides in the the non-template (NT) strand and residues in the σ and β subunits that are formed in the R·P$_o$ state. The color scheme is shown to the right; Chains H and C are in the σ and β subunits respectively. (a2) Same as (a1), except the distances are between different nucleotides and residues (see the box to the right). (b1) Same as (a1) except the nucleotides are in the template strand and the residues are labeled in the box to the right. The channel opening time interval $\Delta\tau_w$ ($\approx 19\mu s$ to $\approx 28\mu s$) is explicitly shown. (b2) Same as (b1) except the nucleotides and residues (see the box to the right) are different. (b3) Same as (b1), except $d_{-2,1031C}$, $d_{-1,1033C}$, $d_{+1,1033C}$, and $d_{+2,1093D}$ are plotted in black, red, green, and blue, respectively. (c) Variation in the distances between nucleotides on the template strand and residues on the RS domain that do not interact in the R·P$_o$ state as a function of t. The box on the right gives the color scheme. (d) R (yellowish green) and S (blue) domain are shown in ribbon representation, and the promoter in the bubble region is highlighted with spheres. The σ and β' subunits are shown in surface representation. (e) The strain energy between DNA and RNAP as a function of time. The three steps in the transcription bubble formation are in red, yellow, and blue shades in (a)-(c) and (e).

Interestingly, the accumulated strain energy and $R(t)$ both reach a maximum (Figs. 5b and 6e) prior to the start of the bending of downstream DNA. Even after the channel closes bending continues till the value of $\Psi(t)$ in the R·P$_o$ state is reached

(Fig. 5c). The complex internal motions show that the excluded volume interactions between certain regions of the template strand the RS domain enable the "gate" of the channel to widen for the duration $\Delta\tau_w$ until the downstream dsDNA can gain access to the active channel.

3.5.8 Structural fluctuations are enhanced in the slow trajectories:

Rapid formation of $R{\cdot}P_o$ resulting in the transcription bubble formation, which begins with local melting of the -10 element followed by DNA scrunching into RNAP and channel opening by the RS mechanism, is the characteristic of fast trajectories. Premature opening of base pairs that are away from the -10 element compromises the efficiency of the $R{\cdot}P_c \rightarrow R{\cdot}P_o$ transition. In this case, the template strand remains outside of RNAP while the non-template strand forms native-like contacts with the RS domain (Fig. 3.14 (a) and (b)). As a result the interactions between the template strand and the RS domain take considerably larger time to develop resulting in the slowing down of the transcription bubble formation. However, the role of the RS domain, which is prominent in the widening of the channel, is essentially the same as in the fast trajectories. After the template strand starts interacting with the RS domain, the sequence of events similar to that shown in Fig. 3.13 (a-d) transpire. Interactions between the template strand sites -2 to 3 and the RS residues Ala349, Val346, Lys167 and Asn187 (Fig. 3.14 (c)) that are absent in the $R{\cdot}P_o$ state serve as the driving force to move the RS domain and widen the RNAP main channel (Fig. 3.14 (f)). Comparison of the fast and slow trajectories shows that the structural

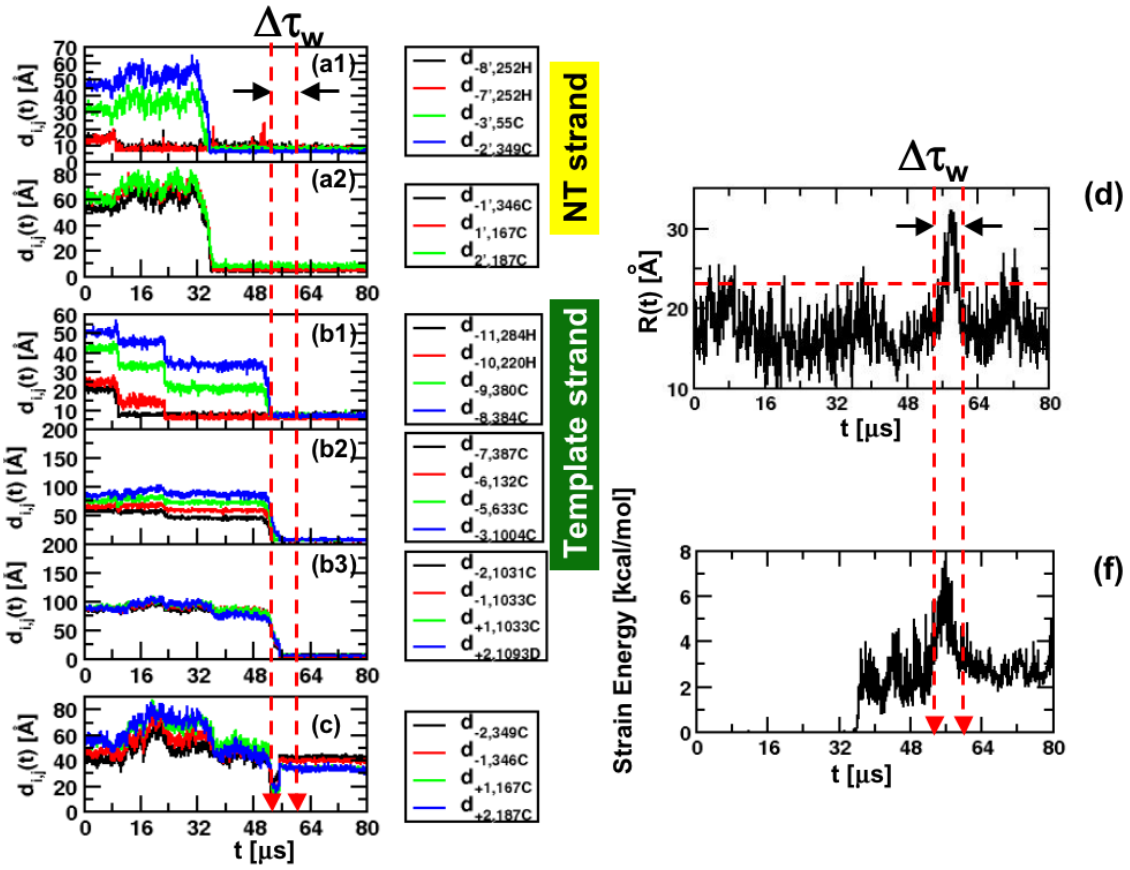


Figure 3.14: (a1) Evolution of the distances in a slow trajectory between nucleotides in the the non-template (NT) strand and residues in the σ and β subunits that are formed in the R·P$_o$ state. The color scheme is shown to the right; Chains H and C are in the σ and β subunits respectively. (a2) Same as (a1), except the distances are between different nucleotides and residues (see the box to the right). (b1) Same as (a1) except the nucleotides are in the template strand and the residues are labeled in the box to the right. The channel opening time interval $\Delta\tau_w$ ($\approx 53\mu s$ to $\approx 60\mu s$) is explicitly shown. (b2) Same as (b1) except the nucleotides and residues (see the box to the right) are different. (b3) Same as (b1), except $d_{-2,1031C}$, $d_{-1,1033C}$, $d_{+1,1033C}$, and $d_{+2,1093D}$ are plotted in black, red, green, and blue, respectively. (c) Variation in the distances between nucleotides on the template strand and residues on the RS domain that do not interact in the R·P$_o$ state as a function of t. The box on the right gives the color scheme. (d) The distance change between the closest points, residue Pro244 and Ala122 on the RS domain and the β' domain as a function of time averaged over ten fast trajectories. Channel expansion occurs at t=$\tau_w \approx 53\mu$s (red arrows) to accommodate downstream dsDNA only. (e) The strain energy between DNA and RNAP as a function of time. The three steps in transcription bubble formation are in red, yellow, and blue shades in (a)-(c) and (e).

fluctuations in the RS domain, as measured by $d_{AB}(t)$ and $\Omega(t)$ are greatly enhanced in the fast trajectories.

3.5.9 Deletion mutation study shows that removing S results in partial formation of the $R \cdot P_o$

It is clear that the RS domain plays an important role in RNAP channel widening process. Stabilizing interactions between the R domain and the promoter DNA facilitate the transcription bubble formation. Excluded volume interactions between the template strand and the RS domain transiently result in movement of the RS domain, which results in the opening of the RNAP main channel by about 7 $\mathring{A}$. After the promoter sequence enters the RNAP main channel, the channel closes to pretect the dsDNA inside it.

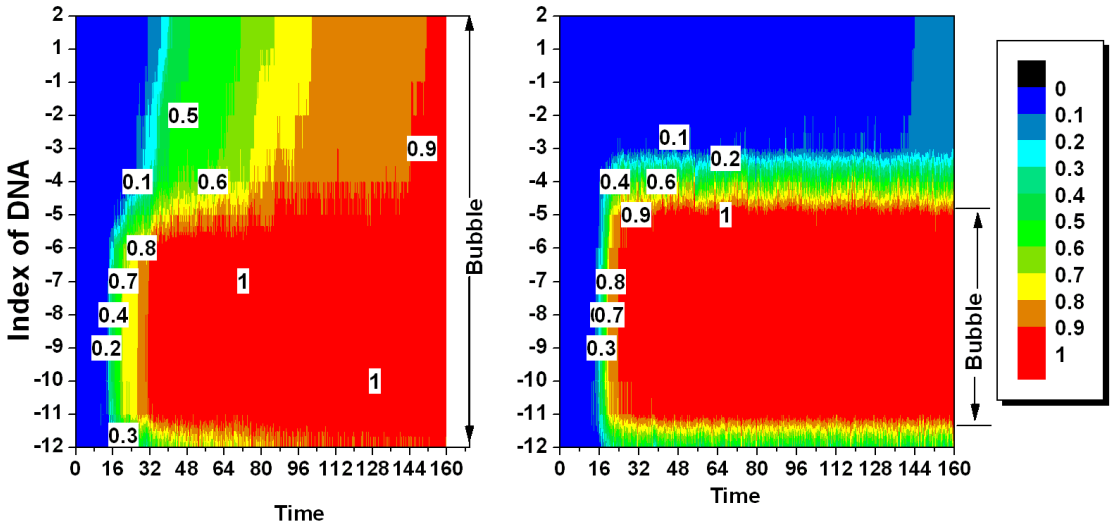


Figure 3.15: Base-pair opening probability for promoter region -12 to +2 as a function a time. WT RNAP on the left, RNAP with deletion mutation of the rope domain on the right. Color bar indicates the base-pair opening probability with red and blue colors show the probability equal to 1 and 0.

To further examine the role of RS domain in the formation of transcription bubble, we perform deletion mutation on RNAP by removing the S domain, and repeat our simulation. To compare the dynamics of the mutant to that of the wild-type RNAP, we defined the base-pair opening probability of the promoter region -12 to +2 as $P(i,t) = \frac{\sum_{M=1}^{N} F(d(t)_{i,i'}^{M} \geq 11\mathring{A})}{M}$, where $d(t)_{i,i'}^{M}$ is the distance between nucleotides i and i' in trajectory M at time t, $F(d(t)$ is a step function which is 1 if $F(d(t) \geq 11\mathring{A})$ and 0 otherwise, and $N = 20$ is the total number of trajectories we collected. In Fig. 3.15, from red to blue, the opening probability decreases from 1 to 0. At $t = 160$, the entire transcription bubble is formed in WT (red colors) while the base pairs opened from -12 to -5 in mutant (Fig.3, right panel). In addition, the opening probability becomes 1 very fast from base pair -10 to -7, and the probability increases from 0 to 1 gradually from -6 to +2, which indicates that dsDNA unzips from -10 to +2.

Comparing to the experimental results by Darst *et al* for the same deletion mutated RNAP, that the transcription bubble formed from -12 to -7 for mutant, our results are reasonable. The partial formed bubble state (-12 to -5) is considered an intermediate state of formation of transcription bubble [111]. Our simulation shows that, without RS domain, the active site channel is open for downstream DNA to enter. After finishing local melting, the promoter DNA enters the active channel directly without further melting to overcome the energy barrier caused by exclusive volume interaction between RS domain and downstream DNA.

3.6 Concluding remarks

Using simulations we have shown that the the transcription-competent R·P_o complex formation, which is known to occur through a series of intermediates [38, 39], occurs in several steps and results in the transcription bubble formation. In fully accommodating the DNA, the internal dynamics of RNAP, which is often hidden in many conventional experiments, plays a crucial role. The proposed three step formation of transcription bubble is similar to the time-resolved hydroxyl radical foot printing experiments [38, 112], which characterized the formation of the complex between the T7A1 promoter and the *E. Coli.* RNA polymerase. From the time dependent changes in the appearance of protection against hydroxyl radical, they surmised that globally there are three steps in the formation of R·P_o. In the first two stages the protection is attributed to interactions of melted (template and non-template strands) with the RNAP, while the slower protection seen in the last stage is due to the entry of the dsDNA (+3 to +21) into RNAP. Our work suggests that the latter process is slow because it involves channel opening that is intimately coupled to the internal dynamics of the RNAP. However, in a different study [39] suggested, using analysis of hydroxyl radical and potassium permanganate ($KMnO_4$) , that the promoter (λ_RP) does not melt from -11 to +2 until after gaining entry into RNAP. The dynamical structural and energetic changes in our simulations do not support this picture. In particular, the stabilization of melted single strand around the -10 element by the aromatic residues of the σ 2.3 could also prevent $KMnO_4$ from reacting with the intermediate [113]. If this was the case then the complex

formation mechanism would be similar to the present and previous [38] studies.

It is worth pointing out a few limitations of the present work. (1) We did not include $\sigma_{1.1}$, which is found to play some role in promoter recognition and accelerating the formation of open complex at some promoters [114]. It is possible that $\sigma_{1.1}$ plays some role in the relaxation step, to help promoter DNA bend in the downstream region outside the bubble [97, 40], but elucidation of its precise role awaits further studies. (2) We have used, out of necessity, coarse-grained description of the kinetics of $R{\cdot}P_o$ formation. The simplification not only underestimates the overall time scale of the $\mathrm{R{\cdot}P_c} \rightarrow \mathrm{R{\cdot}P_o}$ transition but also might lead to an oversimplification of the dynamics.

3.7 Methods

3.7.1 Rationale of the Hamiltonian switching method

We introduce the Hamiltonian switching method to trigger the conformational changes of biopolymers. Our basic assumption is that local strains build up locally in a protein (i.e., caused by ligand binding in active sites) will eventually propagate to the rest of the structure. This method assumption is based on fluctuation dissipation theorem. Consider a constant field f (local strains caused by ligand binding) applied to a protein in equilibrium. If the field is weak, the changes in Hamiltonian is a linear function of the field,

$$< H(x) >_f - < H(x) >_0 = fx \tag{3.2}$$

where $< H(x) >_f$ denotes the Hamiltonian of a protein when the field is applied and $< H(x) >_0$ is the equilibrium value in the absence of the field ([1, 53]). The effect of the field on the protein is expressed by a potential such as $H_{ext}(x) = fx$, where quantity x, i.e. the protein conformation, is conjugate to the field f. In such cases, the relaxation of the protein conformation x in response to a filed f is related to a response function $\mu(t)$ (see Fig.3.16 (a-b))

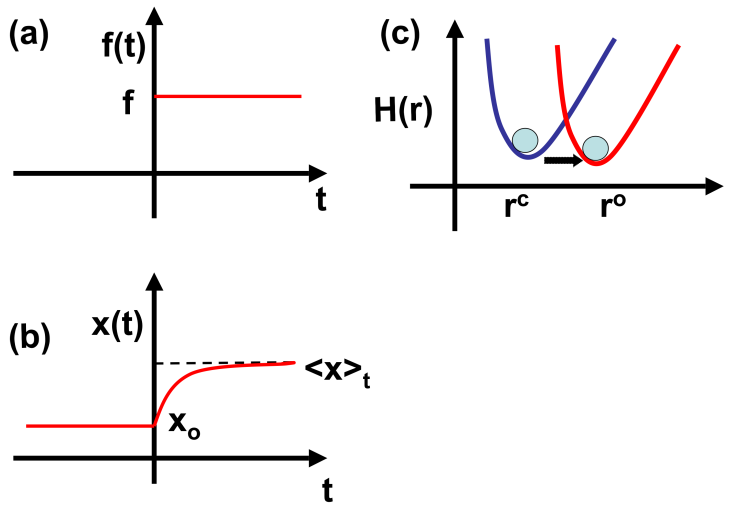


Figure 3.16: Relaxation function. (a) and (b), when a constant external field is switched on at $t = 0$, the average of a physical quantity $x$relaxes from the equilibrium value in the absence of the field $< x >_0$ to that in the presence of the field $< x >_t$. The time dependence for $t > 0$ is described by the relaxation function 3.3. (c) r^c and r^o are conformations for close state (ligand-unbound) state, and open state (ligand-bound state). The corresponding Hamiltonian function is H^c and H^o.

$$< x(t) >_f - < x >_0 = \int_{-\infty}^{t} dt' \mu(t - t') f \tag{3.3}$$

In the context of proteins, the Hamiltonian for a closed (ligand unbound) state is H^c at $t = (-\infty, 0)$. At time t, an external force, $-\frac{\partial H^o}{\partial \vec{r}}$, is added. Here, H^o is the Hamiltonian for the open (ligand bound) state. According to the fluctuation

dissipation theorem, protein conformation will relax from that of the closed state to the open state with some relaxation function (see Fig.3.16 (c)).

3.7.2 Technical details on triggering the conformational changes

It is essential to choose the Hamiltonian switching time τ to be much less than the relaxation time of the conformational changes in order to obtain the correct dynamics. The time scale of building up local strain upon ligand binding, τ^{lig}, is on the orders of ps $\sim$ ns, while the time required for large-scale conformational changes, τ^{conf}, is on the orders of μs $\sim$ ms. These time scales determine that τ must be chosen so that $\tau \sim \tau^{lig} \sim \tau^{conf}$. In a previous study [115], a similar Hamiltonian switching method was used to study the rotary moter F1-ATPaseis, in which the forces driving the conformational transition is induced in a single time step. In our simulation of the transcription bubble formation, we used a linear formula to implement the changes in the forces that drive the R·P_c $\rightarrow$ R·P_o transition. In our procedure the change in the dynamics is implemented using

$$r_{ij}^{c\rightarrow o} = \frac{(K-k)\cdot r_{ij}^{c} + k \cdot r_{ij}^{o}}{K}, \tag{3.4}$$

where, $K = 100$, and k is increased by 1 every 1000 steps from 1 up to 100. The value $K \cdot 1000 = 100,000$. Since we use smaller time step, $h = 0.016$, to maintain the simulation stability during the Hamiltonian switching period, the estimated switching time corresponding to $0.08\mu s$. The corresponding Hamiltonian and forces during Hamiltonian switching can be easily calculated from $r_{ij}^{c\rightarrow o}$ and the energy function Eq.1.12.

Using a linear equation and a value of $K > 1$ not only ensures that there is a lag time between ligand binding and the associated response, but also eliminates computational instabilities in the distances between certain residues that change dramatically during the transition. The "loading" rate can be altered by varying the number of steps used to switch from r_{ij}^{c} to r_{ij}^{o}.

Equation 3.4 describes a simple framework for transforming a protein between specific conformational states, which does not inherently imply any structural constraints in their usage with the Hamiltonian studied in this work. More generally, we can use a exponential function to trigger conformational changes of a protein. A simple form of the exponential function can be written as

$$r_{ij}^{c\to o} = e^{-t/\tau} r_{ij}^{c} + (1 - e^{-t/\tau}) r_{ij}^{o}, \tag{3.5}$$

where τ is the characteristic time during which the forces from $\mathrm{R}\cdot\mathrm{P_o}$ drives $\mathrm{R}\cdot\mathrm{P_c} \rightarrow \mathrm{R}\cdot\mathrm{P_o}$ transition. From $t = 0$ to ∞, r_{ij} changes exponentially from r_{ij}^{c} to r_{ij}^{o}. By varying τ, we can easily assess the effect of τ on the relaxation dynamics of conformational changes, which can be monitored using τ^{conf}. In the following, we choose RNAP as an example, and test this approach for the transcription bubble formation. We varies τ from $10,000$ steps to $320,000$ steps, which corresponds to $0.08\mu s$ to $2.56\mu s$ in real time (see Fig. 3.17).

We monitor the kinetics of transcription bubble formation as a function of τ, using $P_u^{\tau}(t)$, which is the function of trajectories that reach the target conformation at time t [116, 55]:

$$P_u^{\tau}(t) = \frac{1}{M} \sum_{i=1}^{M} \delta(t - \tau_{1,i}) \tag{3.6}$$

where, $\tau_{1,i}$ denotes the first passage time for the i^{th} trajectory, i.e. the time when a trajectory adopts the target conformation for the first time. The mean first passage time τ_{MFPT} to the target conformation gives the characteristic conformational transition time τ^{conf}. In our simulations, we find that $P_u^{\tau}(t)$ can be well fit with

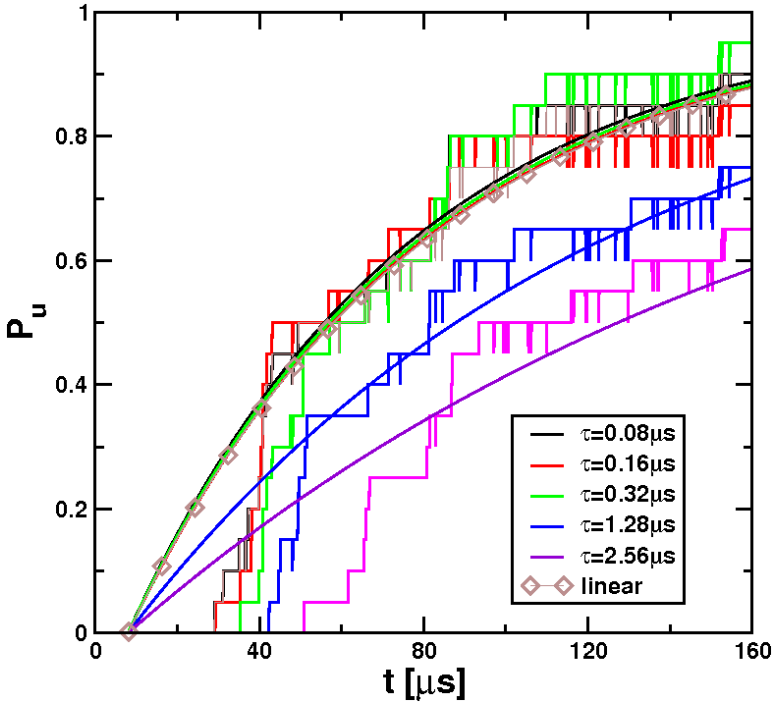


Figure 3.17:
Probability of successful switching as a function of time, P_u, at various switching condition. $\tau = 0.08\mu s$, $0.16\mu s$, $0.32\mu s$, $1.28\mu s$, and $\tau\mu s$ are plotted with black, red, green, blue, and yellow lines. Linear switching is plotted in brown for reference. The exponential fitting line for each curve is plotted using black, red, green, blue and purple lines for $\tau = 0.08\mu s$, $0.16\mu s$, $0.32\mu s$, $1.28\mu s$, and $2.56\mu s$. Squares are shown for the fitting line to the linear switching curve.

single exponential of the form (see Fig: 3.17):

$$P_u^{\tau}(t) = 1 - exp(-t/\tau^{conf}) \tag{3.7}$$

For $\tau = 0.08\mu s$, $\tau = 0.16\mu$st, and $\tau = 0.32\mu s$, we fit P_u with $\tau_{conf} \approx 45\mu s$, whereas for $\tau = 1.28\mu s$ and $\tau = 2.56\mu s$, we fit τ_{conf} with $70\mu s$ and $107.5\mu s$. As

a reference, we fit $P_u(t)$ obtained using linear switching (Eq: 3.4), and get $\tau_{conf} \sim 45\mu s$. The comparison of τ_{conf} at various τ shows that the upper limit of τ^{max} is $< 1\mu s$, beyond that, the artifacts of the switching method affect the kinetics of the conformational transition.

3.7.3 Persistence length of DNA

Persistence length, l_p, indicates the stiffness of semiflexible polymers. To ensure that the SOP model for DNA is reasonable, we calculate l_p by fitting the end-to-end distance distribution, $P(R)$, to a simple analytical equation given by Thirumalai *et al* in [117],

$$P(r;t) = 4\Pi C \frac{r^2}{(1-r^2)^{9/2}} exp[-\frac{3t}{4}\frac{1}{(1-r^2)}] \tag{3.8}$$

where, $r = \frac{R}{L}$, and $t = \frac{3L}{2l_p}$. R represents end-to-end distance, $L = N \cdot a$ is the contour length of a polymer that has N segment of size a, and C is a constant.

Using SOP model, we carried out three sets of Brownian dynamics simulations for 1) A free 75 base-pair double stranded B-form DNA. 2) 75 base-pair DNA in R·P_c complex. 3) 75 base-pair DNA in R·P_o complex. After equilibration, we calculated $P(R)$ for DNA in three sets (see Fig. 3.18). For the DNA in our simulation, $N = 75$, $a = 3.4\mathring{A}$, and $L = 75 \times 3.4 = 255\mathring{A}$. With these numbers, we fit $P(R)$ of the free DNA to Eq. 3.8 with $l_p = 44.3nm$, which agreed well with the experimental value, 46-50 nm, of DNA in aqueous solution. Compared to free form, $P(R)$ for DNA in R·P_c becomes wider and shifts to the left due to interactions with RNAP. The structure of RNAP complex with fork-junction DNA shows that promoter DNA

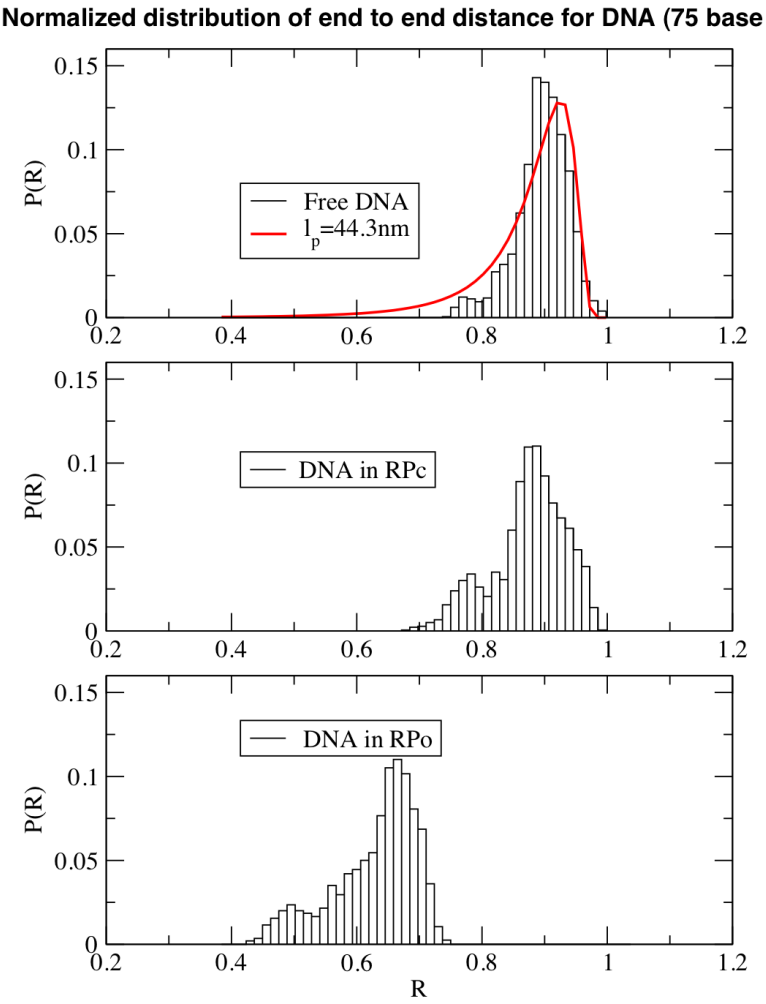


Figure 3.18: Distribution of end to end distances: (a) Free DNA, (b) DNA in R·Pc (c) DNA in R·Po

bends by about 36^o around -35 element. For DNA in R·P_o, the distribution $P(R)$ shifts even more to the left as a result of DNA melting and bending. SOP model, although ignores many complex interactions in DNA (hydrogen bonding, stacking, electrostatic interaction, eta), captures the basic polymer properties for DNA, and can produce meaningful dynamics for DNA.

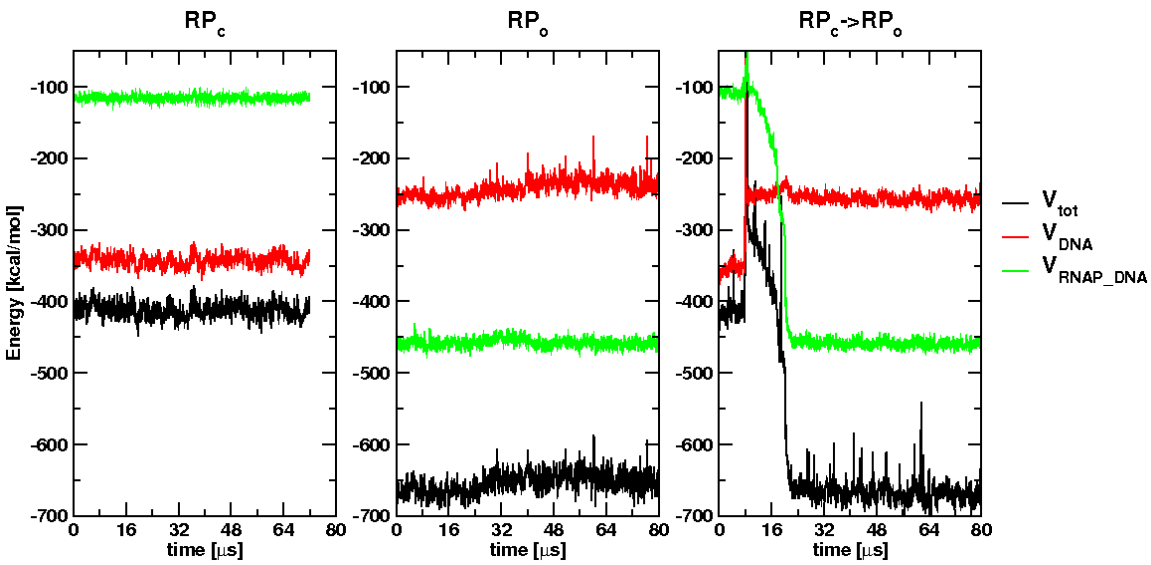


Figure 3.19: Energy of RNAP-DNA complex at function of time. From left to right, energy in R·P$_c$, energy in R·P$_o$, energy changes in a fast trajectory, and energy changes in a slow trajectory. The total energy V_{tot} is plotted in black, energy for DNA V_{DNA} is in read, and energy between DNA and RNAP V_{prot_DNA} is in green.

3.7.4 Energy changes of RNAP

Using RNAP as an example, we also calculated the energy difference between the two states R·P$_c$ and R·P$_o$ (see Fig. 3.19). The total potential energy, V_{tot}, for R·P$_c$ is $\sim -400kcal/mol$, and for R·P$_o$ is $\sim -650kcal/mol$. $rmR{\cdot}P_o$ has lower energy and is more energetically favorable. The potential energy is further divided to energy for DNA V_{DNA} and that between the RNAP and DNA $V_{prot-DNA}$. V_{DNA} increases from $\sim -350kcal/mol$ in R·P$_c$ to $\sim -250kcal/mol$ in R·P$_o$ due to loss of contacts in the bubble region. However, $V_{prot-DNA}$ decreases from $\sim -120kcal/mol$ to $\sim -450kcal/mol$ in the R·P$_c$ · R·P$_o$ transition due to stabilization of single stranded DNA in the bubble region (position -10 to +2) by RNAP. This stabilizing interaction is so strong that V_{tot} in R·P$_o$ is lowered even thought V_{DNA} is increased, which

essentially decides DNA spontaneous melting and transcription bubble formation.

3.7.5 Free energy profile

To calculate free energy, we choose size of transcription bubble as a reaction coordinate, which is defined as the average distance between the base pairs in the bubble region $R = \frac{\sum d_i, i'}{N}$, where i and i' are complementary base pairs in the bubble region, d_i, i' is the distance between them, and N is the number of base pairs in the bubble (N =14 in our calculation). We carried out Brownian dynamics simulation in R·P$_c$ and R·P$_o$ for $t = 160\mu s$ in overdamped regime. In both case, R fluctuates as a function of time. The distribution of R, $P(R)$, is calculated for R·P$_c$ and R·P$_o$, which can fit to Gaussian function with average values $< R >= 11\mathring{A}$ and $37.4\mathring{A}$, respectively. The width of the distributions $P(R)$, σ_R is also obtained, which is $0.2\mathring{A}$ for R·P$_c$ and $0.5\mathring{A}$ for R·P$_o$. The larger σ_R in R·P$_o$ state indicates that promoter DNA becomes more flexible in R·P$_o$ because of bending and melting of dsDNA. From $P(R)$, the free energy profile can be calculated using a simple equation, $F(R) = -k_B T log(P(R))$, which shows minima around $< R >$ (see Fig. 3.20, [118]).

For the R·P$_c$ $\rightarrow$ R·P$_o$ transition, we generated 30 trajectories to sample the conformational space. We calculated the bubble size distribution $P(R)$, and free energy as function of R similarly (Fig. 3.20 (c)). R increases from $11\mathring{A}$ to $37.4\mathring{A}$ during the transition. Intermediate states, corresponding to the partial formed bubbles, exist at $R = 18\mathring{A}$ and $R = 25\mathring{A}$. The frustrated free energy profile reflects

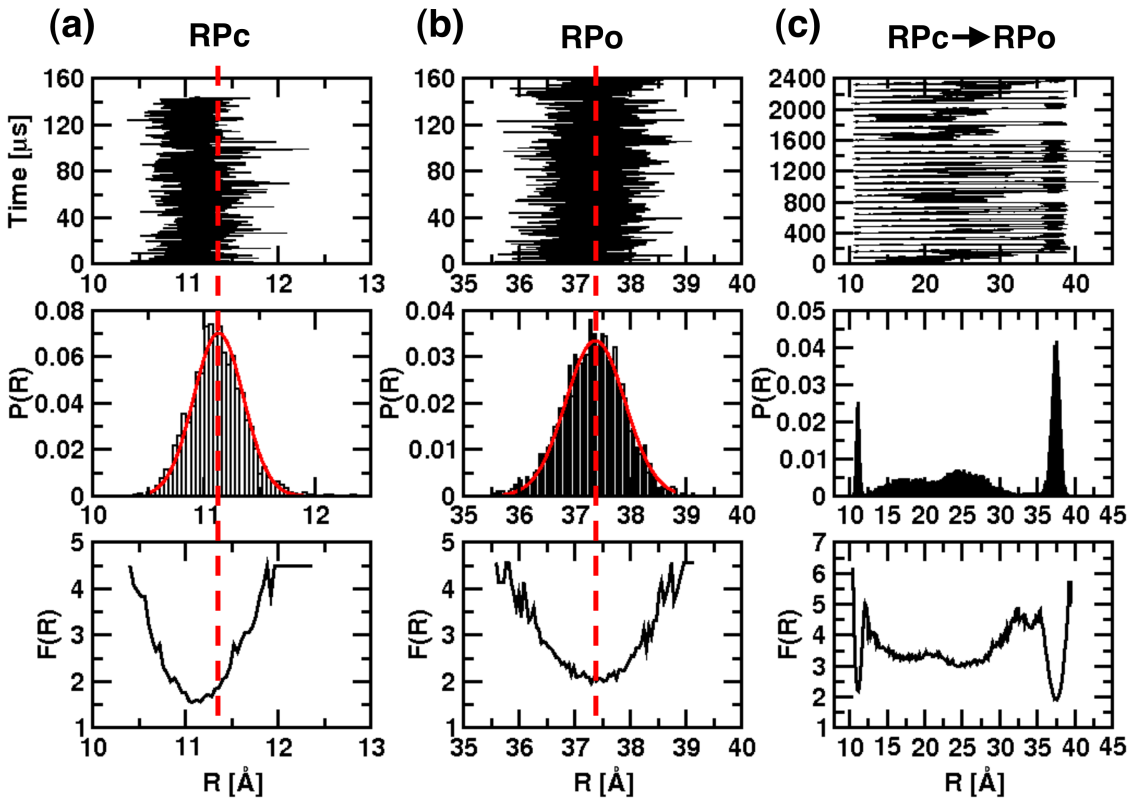


Figure 3.20: Free energy as a function of bubble size R: (a) For RNAP-DNA complex in R·P$_c$ state, R as a function of time, $P(R)$ and $F(R)$ are plotted from top to bottom. Red line on $P(R)$ plot is the Gaussian fitting. (b) Same as (a), except these are for R·P$_o$ state. (c) Same as (a), except these are for transition from R·P$_c$ to R·P$_o$. The most probable bubble size is labeled using red dashed lines, which is $\sim 11\mathring{A}$ in R·P$_c$, and $\sim 37.4\mathring{A}$ in R·P$_o$.

the stochastic property of the transition R·P$_c$ $\rightarrow$ R·P$_o$.

Chapter 4

Eukaryote RNA polymerase II and the conformational changes involved in transcription

4.1 Overview

RNA polymerase II (pol II) is responsible for synthesis of eukaryotic mRNAs in the course of gene transcription. In broad terms, the transcription process involves three main steps: 1, promoter binding and formation of the transcription initiation complex [119, 120, 121]. 2, synthesis of nascent RNA and transcription elongation[122, 123, 124, 125]. 3, the cessation of nascent RNA molecule and termination of transcription[126, 127]. Among the three steps in the transcription cycle, initiation is perhaps the most complicated since it involves transcription factors assembly, recognition of the promoter DNA, unwinding of the DNA double strand, and formation of a stable transcription bubble inside the active-site cleft of Pol II where synthesis is carried out. All of these functions are accomplished through participation of multiple general transcription factors, for example, TFIIB, -D, -E, -F and -H, which bind to Pol II during transcription[119, 120]. Binding of multiple transcription factors makes the transcription initiation in eukaryotic Pol II considerably more complicated than the corresponding processes in bacteria. Elongation and termination steps are relatively simple since the general transcription factors

have dissociated from Pol II in these stages. Therefore, the transition from initiation to elongation does not involve as many intermediate states as in the initiation process.

Kornberg and coworkers made several important discoveries in not only solving the structures of Pol II, but also in the elucidation of the biophysics of transcription [128, 129, 130]. The crystal structures of Pol II holoenzymes and complexes with promoter DNA have been solved. Comparing the two forms of Pol II holoenzyme[128], four mobile modules are recognized (Fig.4.2 (a)), among which the clamp module is found to undergo a large swinging motion, which serves as a multifunctional element in transcription[128, 131]. For the structure of a Pol II elongation complex with DNA-RNA hybrid, closure of the clamp and ordering of a series of "switches" in the active center are observed [129].

Despite the availability of structures of Pol II in varies conformation states, the dynamics pathways that Pol II undergoes from one conformational state to another in the transcription cycle remains unknown[132, 130]. Here, in order to study the intrinsic motions of Pol II, we employed normal mode analysis (NMA) and structural perturbation method (SPM) to identify the function-related normal modes and residues that transmit the signals for dynamical transition from one state to another. In addition, we performed Brownian dynamics simulations using a self-organized polymer (SOP) model[35], to simulate the F1→F2 and the F2→EC transitions of Pol II. Dynamics of the global motions of the three mobile modules are analyzed, among which the clamp dynamics are found to dominate the intrinsic dynamics of Pol II. The formation and rupture dynamics of multiple native contacts

and salt bridges that trigger the global motions are also obtained. From both NMA and Brownian dynamics simulation, we discovered that the open-close and back-forth motions of the mobile clamp in F1→F2 transition accompany the transcription initiation function of Pol II, while the open-close and outside-in motions of the mobile clamp in F2→EC transition accompany the transcription elongation.

4.2 Architecture of Pol II

Pol II comprises 12 subunits, numbered Rpb1 to Rpb12, with a total mass of $\sim$ 500kDa (Fig.4.1). Among the 12 subunits, Rpb4 and Rpb7 are dispensable for transcription, therefore are not included in the 10-subunit core enzyme of Pol II. The crystal structures show that Pol II adopts a crab-claw shape just like RNAP (Fig. 3.10), with the two largest subunits, Rpb1 and RPb2, being the two pincers of the "claw". Pol II is overall negatively charged, but between Rpb1 and RPb2, there is a positively charged active-site cleft that forms the DNA binding channel. Active sites locate in the floor of the active-site cleft, and are marked by a "Mg^{2+}" ion. It is interesting that DNA path is block by a protein "wall" in the back of the active-site cleft, which requires DNA to bend at nearly right angles to the direction of the incoming DNA in the active-site cleft (Fig. 4.1 (b)).

Two crystal forms of Pol II core enzyme, form 1 (F1) and form 2 (F2), have been derived by Kornberg and coworkers [128]. By comparing the structures of the F1 and F2, they further divided the 10-subunit Pol II into four mobile modules, i.e. clamp, jaw-lobe, shelf and core (Fig: 4.2 (a) and (d)). The clamp module comprises

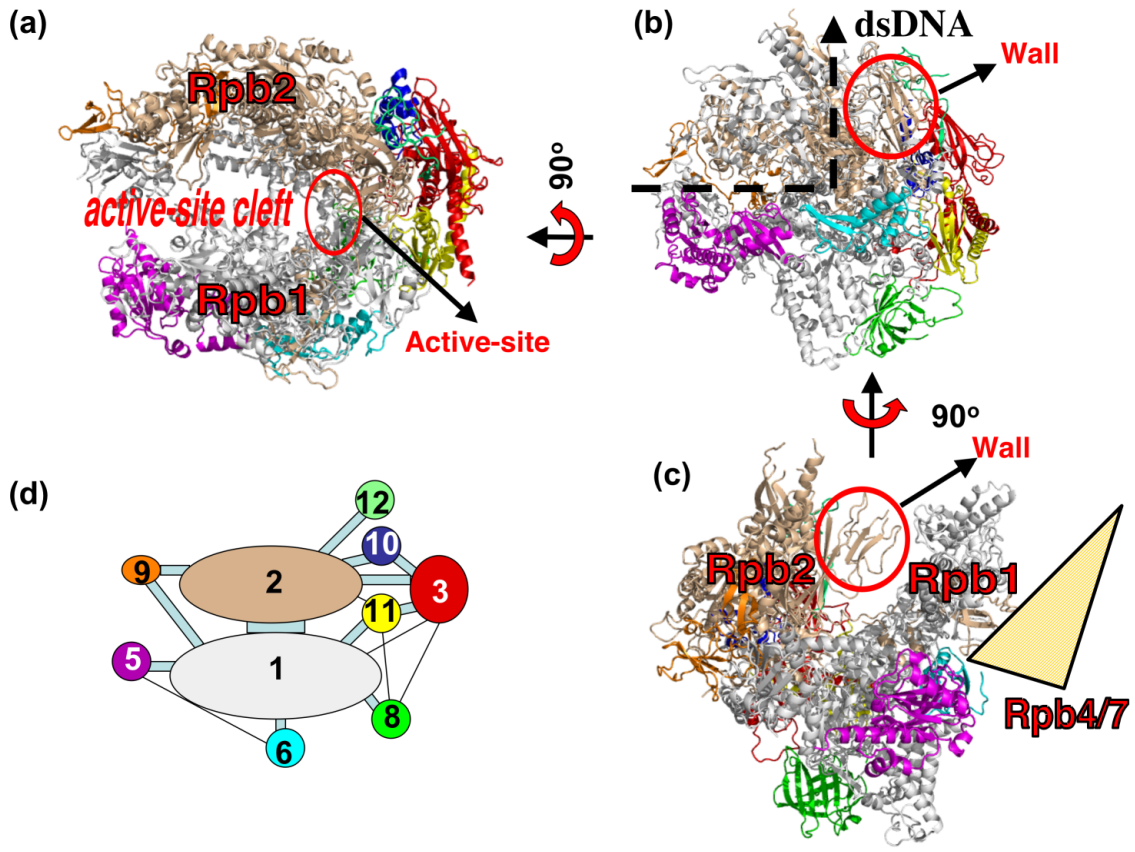
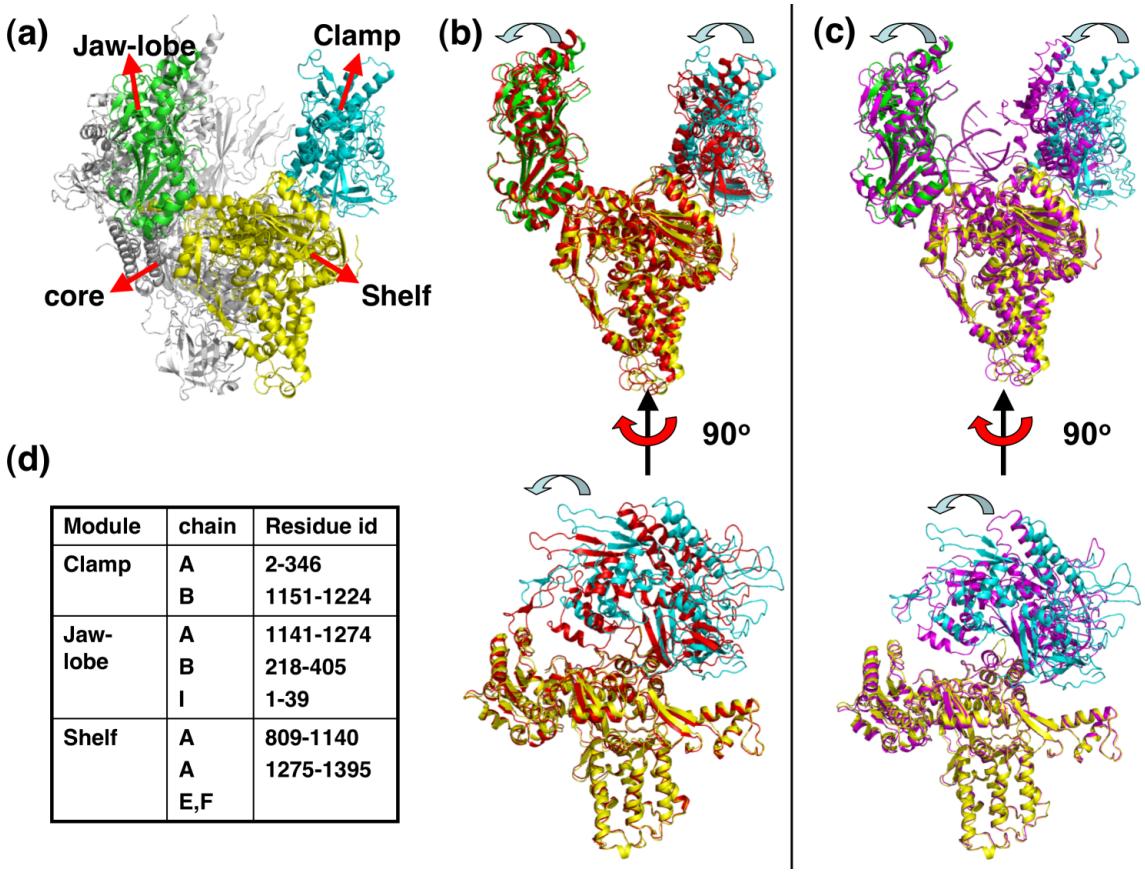


Figure 4.1: Crystal structures of pol II (PDB entry: 1I3Q). (a) Top view of the 10-subunit Pol II. Rpb1 and Rpb2 are labeled, between which is the active-site cleft. The active-site is indicated with a circle and labeled. (b) Side view of the same structure. "Wall" is circled and labeled. Dashed line shows DNA bending about 90^o near the "wall" of Pol II. (c) Front view of the same structure. Rpb1, Rpb2, and "Wall" is indicated and labeled. The yellow triangle represents the additional subunits Rpb4/7. (d) The organization of the 10 subunits in top view. Each oval represents a subunit and is colored using same scheme as in the structure.

the N-terminal region of Rpb1 (residues 2 to 346) and the C-terminal region of Rpb2 (residues 1151-1224), and is the most mobile module of the four [133]. The jaw-lobe module comprises the jaw (residue 1141-1274 of Rpb1 and residues 1-39 of Rpb9) and the lobe (residues 218-405 of Rpb2) domains. The shelf module contains part of Rpb1 (residues 809-1140, residues 1275-1395), Rpb5, and Rpb6. The jaw-lobe and shelf modules are less mobile than the clamp. The rest of the structure belongs to the



Module	chain	Residue id
Clamp	A B	2-346 1151-1224
Jaw-lobe	A B I	1141-1274 218-405 1-39
Shelf	A A E,F	809-1140 1275-1395

Figure 4.2: Four mobile modules of pol II. (a) Pol II is shown in front view, clamp, jaw-lobe, shelf, and core modules are colored cyan, green, yellow, and grey, respectively. (b) Two forms of Pol II (F1 and F2) are overlapped to show the structural changes. On the top, front view is shown with core modules removed for clarity. The three modules, clamp, jaw-lobe, and shelf in F1 state are colored cyan, green, and yellow as in (a), while in F2 they are all colored red. The rotation of the clamp and jaw-lobe modules are indicated with grey arrows. On the bottom, side view of Pol II is shown with core and jaw-lobe modules removed. The rotation of clamp module is indicated with a grey arrow. (c) Same as (b) except we overlap Pol II structures in F1 onto EC. EC structure is shown in purple and the structural changes are indicated with arrows. (d) The division of Pol II into four modules.

core module, which occupies half of the mass of Pol II, and is relatively unchanged in the two forms of Pol II. By aligning the structures of the core modules in the F1 and F2, it is shown that the largest structural changes are the clamp rotation about an axis perpendicular to the paper plan of Fig. 4.2 (b, bottom panel). This rotation

of the clamp away from the back wall and closer to the shelf module. In addition, the jaw-lobe module rotates away from the clamp and widens the active-site cleft between the clamp and the jaw-lobe modules for the entry of dsDNA [128].

The F1 and F2 of Pol II core enzyme provide insights into the transcription initiation of Pol II, while the determination of the Pol II elongation complex (EC) structure expends our knowledge of the transcription to the elongation stage [129]. Comparing structures of the F1 and EC based on the four mobile modules shows that the clamp are rotating into the active-site cleft and make the active-site cleft narrower in the EC. The clamp module rotates closer to the shelf module. These structural changes may due to the presence of the DNA-RNA hybrid. In addition, the jaw-lobe domain rotates slightly away from the active-site cleft (Fig. 4.2 (c, bottom panel)).

4.3 NMA analysis: Pol II holoenzymes and TEC

The structures of the F1, F2 and EC illustrate the possible intrinsic motions of Pol II relevant to the transcription function. However, we need to obtain the dynamic pathways connecting these functional states. Here, we analyze the dynamics of Pol II near the attraction basins of the F1 and F2 using NMA and SPM. First, we derive the function-related normal modes of the F1 and F2, then obtain the important structural units involved in the dynamics of Pol II with SPM.

4.3.1 Three intrinsic motions trigger the conformational transitions of Pol II.

We performed NMA for Pol II in the F1 (F2), and extracted the eigenvectors and eigenvalues of the normal modes. The low frequency modes, which are shown to correlate with function, are overlapped to the conformational changes of Pol II in the transition F1→F2 (F2→EC) to identify the normal modes that are related to the structural changes during the transitions F1→F2 and F2→EC. The largest values of the overlap are found for mode 1 and mode 2 of the F1 (see Fig. 4.3 (a)), and mode 1 and mode 3 of the F2 (Fig. 4.3 (b)). Interestingly, mode 1 of the F1 and mode 1 of the F2 are essentially the same eigen-mode which describes an open-close motion of Pol II. The width of the active DNA binding channel formed by the clamp module on one side and the Jaw-lobe module on the other side increases or decreases in this motion (see Fig. 4.3 (c)). In a study for bacterial RNA polymerase, it is found that diameter of dsDNA (22 $\mathring{A}$) is wider than the width of the active DNA binding channel (19 $\mathring{A}$), and the channel opens to allow dsDNA to enter during the transcription bubble formation [41, 92]. However, for the complete 12-subunit Pol II, the heterodimer Rpb4/7 is like a wedge that locates under the clamp module and restricts opening motion of the clamp (Fig. 4.1 (c)). The requirement of the Rpb4/7 in the transcription initiation rules out the possibility of the clamp opening, hence there must be some other mechanisms that facilitate the transcription initiation of Pol II [134, 135].

Mode 2 of the F1 describes the clamp motion close to or far away from the

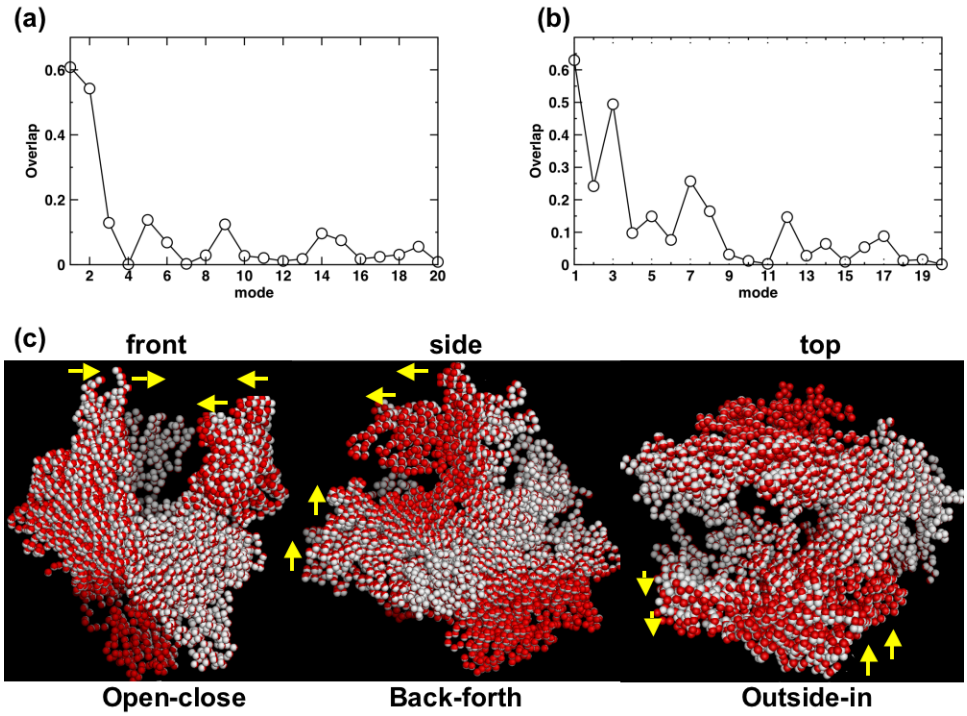


Figure 4.3: (a) Overlaps of the 20 lowest frequency modes with the conformational changes in F1→F2 calculated with Eq: 1.10. Mode 1 and mode 2 are the two modes that overlap best with the conformational changes. (b) Overlaps of the 20 lowest frequency modes with the conformational changes in F2→EC calculated with Eq: 1.11. Mode 1 and mode 3 are the two modes that overlap most with the conformational changes. (c) From left to right, demonstration of open-close motion in front view, back-forth motion in side view, and outside-in motion in top view, respectively. Cα atoms of Pol II is shown with spheres. The open state is in white and close state is in red, arrows are used to indicate the direction of motion.

wall (referred to as the back-forth motion, see Fig. 4.3 (c)). The back-forth motion of clamp affects the DNA by pushing DNA into the active-site cleft to help Pol II to translocate along DNA. The wedge-like structure of Rpb4/7 under the clamp in the complete 12-subunit Pol II will not affect the back-forth motion of the clamp module, which implicates that the back-forth motion plays an important role in the transcription initiation.

Mode 3 of the F2 features the rotation of the clamp around the axis perpen-

dicular to the paper plan of the top view of Pol II (referred to as the outside-in motion, see Fig. 4.3 (c)). In the F2→EC transition, the open-close motion helps to hold DNA duplex inside the active-site cleft during the nucleoside triphosphate (NTP) addition process while releasing it when DNA translocates [128, 136]. The clamp head moves towards (or "in") or away from(or "out") the active-site cleft in the reverse transition. It is believed that Pol II utilizes the outside-in motion to "screw" the downstream DNA duplex towards the active center [128].

4.3.2 Function-related structural units that promotes intrinsic motions are identified using the structural perturbation method (SPM).

Structural perturbation method (SPM) probes the response of the protein to a perturbation at each of the residues (see section 1.3). Here, we used SPM for Pol II and calculated the frequency changes of the three function-related normal modes upon perturbations (see Fig. 4.4 (a)-(c)). For open-close motion, large responses are observed upon perturbing the bridge helix, active sites, switch 1, switch 3, switch 4, and switch 5, which are all known to be important function-related sites. The importance of the bridge helix, whose sequence is highly conserved, can be explained based on its location in Pol II. The long bridge helix connects the two sides of the Pol II "active-site cleft" by directly contacting the jaw-lobe on one side and the joint of the clamp and shelf modules on the other side. The open-close motion of Pol II active-site cleft coordinates with the elongation or shrinkage of the bridge

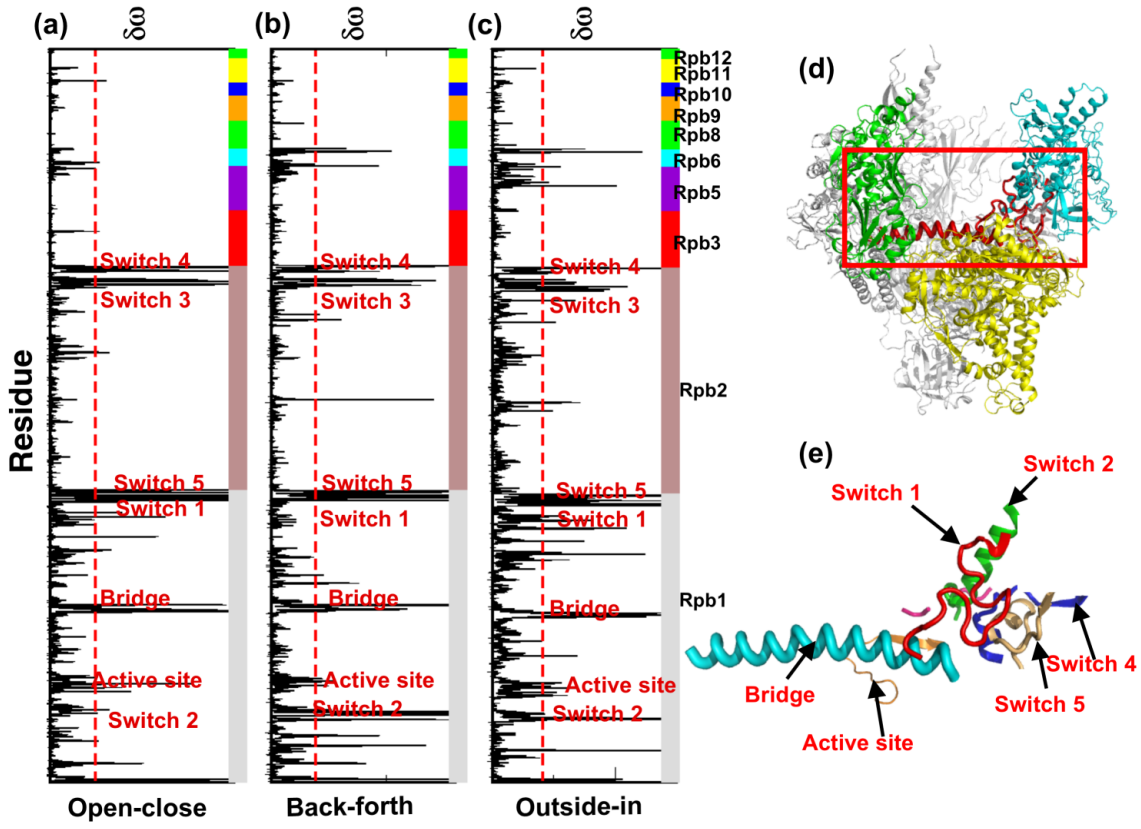


Figure 4.4: Frequency changes of function-related normal mode, $\delta\omega$, as a function of residue index. (a-c) SPM results are shown for open-close, back-forth, and outside-in motions, respectively. On the right of each panel, the division of the subunits is shown and labeled. Red dashed line shows the cutoff value, $2 < \delta\omega >$. Residues that have large $\delta\omega$ are labeled and colored red on (d). (e) Zoom in of the red box in (d), structures of switch 1, switch 2, switch 4, switch 5, active site, and bridge helix are shown(d) Front view of Pol II,

helix. Hence, perturbing the bridge helix affects the communication of the jaw-lobe and clamp modules, and the open-close motion of Pol II. It has been shown that the bridge helix in Pol II is straight but bent and partially unfolded in bacterial RNAP [87]. Since the bridge helix directly contacts the end of the DNA-RNA hybrid in the Pol II elongation complex [129], it is speculated that the bent to straight transition of the bridge helix contributes to the translocation of Pol II along DNA.

At the active sites, a Mg^{2+} ion is persistently bound by three invariant aspartates D481, D483, and D485 of Rpb1, and catalyzes the growing of RNA. Since active site is close to the bridge helix, we think perturbing residues in active site may affect the open-close motion through its interaction with the bridge helix. We also identify multiple switches, switch 1, switch 3, switch 4 and switch 5, which are known to undergo conformational changes and folding and unfolding transitions during Pol II transition from the free form to the elongation complex (Fig. 4.4 (e)) [129]. Among the switches, switch 1, switch 4, and switch 5 forms the hinge region that mediates rotation of the clamp related to the opening and closing of the active-site cleft [129]. It is also shown in a genetic study that the switches are target for inhibitors [137].

For the back-forth and outside-in motions, large responses are observed in all of the sites that are identified in the open-close motion except the active sites and the switch 1. For the active sites, considering the restriction of the open-close motion in the transition initiation process by Rpb4/7, their importance is weakened in the initiation related transition F1→F2. This is reasonable because the important conformational changes are the promoter melting and entry of dsDNA into the Pol II the active-site cleft, which does not involve interactions with the active site. In addition, the switch 2, which is not identified in the open-close motion, is shown to cause large changes on the back-forth motion upon perturbation. Since the switch 2 is a helix with the long axis almost perpendicular to the axis of the bridge helix, it does not affect the close-open motion of the clamp.

Involvement of the structural units, active sites, bridge helix and all of the

switches, indicate a coupling between RNA synthetic reaction and the bridge helix mediated translocations of DNA in the transcription elongation.

4.4 Brownian dynamics simulation of the conformational transitions

4.4.1 The global motion of Pol II is dominated by the motion of the clamp module throughout the transitions.

Brownian dynamics simulation are carried out for the 10-subunit Pol II to study the dynamic pathways connecting conformational states F1 and F2, and F2 and EC. We obtained 37 independent trajectories for the transitions F1→F2 and F2→EC. The global motions of Pol II in these transitions are described by the global root mean square deviation (Δ_G, where G refers to global) of the Pol II conformation with respect to the F1 state, Δ_G^{F1}, and to the F2 state, Δ_G^{F2} (see Fig. 4.5 (a)). Pol II is initially in the F1 basin of attraction with $\Delta_G^{F1} \approx 3\mathring{A}$ and $\Delta_G^{F2} \approx 4\mathring{A}$. At $t \sim 8\mu$s, the transition from F1→F2 is triggered. At $t \sim 10\mu$s, Δ_G^{F1} and Δ_G^{F2} crosses each other, meaning the Pol II conformation leaves the basin of attraction corresponding to F1 and enters the basin of F2. The transition time, τ^G, is defined as the time when $|\Delta_G^{F1}(t) - \Delta_G^{F2}(t)| < 0.1\mathring{A}$ is satisfied. The fully equilibration of Pol II in F2 is reached at $t \sim 32\mu$s with Δ_G^{F1} and Δ_G^{F2} approaches $4\mathring{A}$ and $3\mathring{A}$, respectively.

The local motions of the Pol II are similarly assessed using local RMSD (Δ_L). Knowing that the position of the core module remain unchanged in the transition F1→F2[128], we only analyzed Δ_L of the rest three mobile modules, clamp, jaw-lobe, and shelf (see Fig. 4.5 (b)-(d)). Δ_L for the clamp resembles that of the Δ_Gs,

with Δ_L^{F1} increases from $4\mathring{A}$ in the basin of attraction of F1, to $7\mathring{A}$ in the basin of F2, and Δ_L^{F2} decreases from $7\mathring{A}$ in the basin of F1, to $4\mathring{A}$ in the basin of F2. Comparing to the Δ_Ls for the clamp module, Δ_Ls of the jaw-lobe and shelf modules have much smaller changes. For the jaw-lobe module, the Δ_L^{F1} increases only $1\mathring{A}$ from $4\mathring{A}$ to $5\mathring{A}$ while the Δ_L^{F2} remains to be $4\mathring{A}$. For the shelf module, Δ_L changes about $0.5\mathring{A}$. Comparisons of the Δ_G and Δ_Ls show that the clamp motion dominates the Pol II global motion.

In addition, the local transition time for three modules are similarly derived to be $\tau^L \sim 12\mu$s, 9μs, and 10μs for clamp, jaw-lobe, and shelf modules, respectively. Comparing with the global transition time, $\tau \sim 10\mu$s, the local transition of the jaw-lobe module is finished first, which is followed by the shelf module. The motion of the clamp module finishes last in the entire conformational transition, meaning the clamp motion limits the speed of the F1→F2 transition.

The F2→EC transition is similarly analyzed (see Fig.4.5 (e)-(f)), and the RMSD curves show that the clamp module continues to dominate the transition. Large motions of the clamp module are found to play dominant role in the Pol II conformational transitions and its functions.

4.4.2 The clamp motion results in narrowing/widening of the DNA channel in F1→F2 (F2→EC) transition.

To study the motion of the clamp module in detail, we defined the centers of mass of the jaw-lobe, clamp, and shelf modules to be J(t), C(t) and S(t), respectively

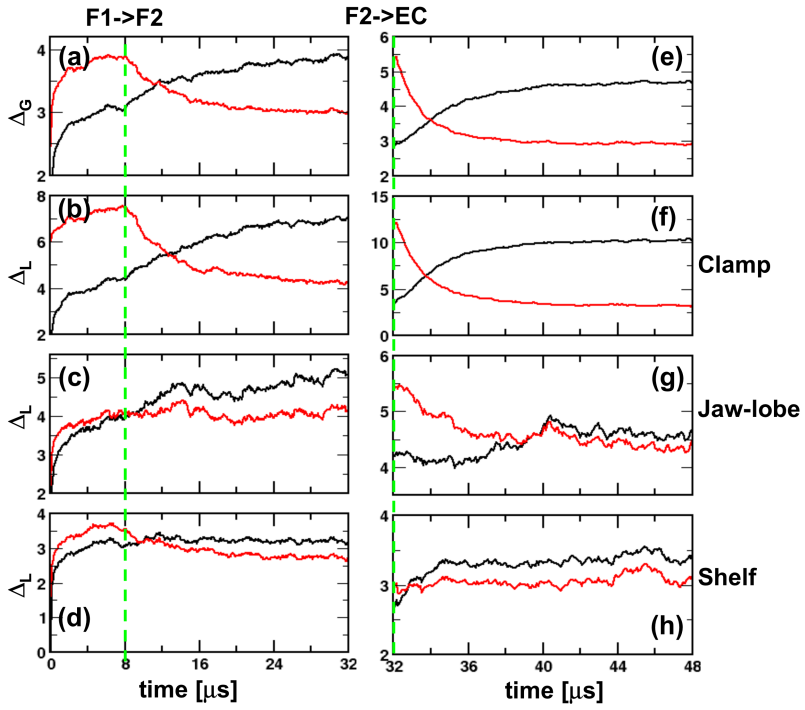


Figure 4.5: RMSD as a function of time. (a) Δ_G at a function of time for the F1→F2 transition. Black and red curves are for Δ_G^{F1} and Δ_G^{F2}, respectively. Green dashed line shows the time when the hamiltonian is switched. (b) same as (a) except these curves are for Δ_L for the clamp module. (c) same as (a) except these curves are for Δ_L for the jaw-lobe module. (d) same as (a) except these curves are for Δ_L for the shelf module. (e) same as (a) except these curves are for Δ_G in the F2→EC transition. (f) same as (b) except these curves are for Δ_G in the F2→EC transition. (g) same as (c) except these curves are for Δ_G in the F2→EC transition. (h) same as (d) except these curves are for Δ_G in the F2→EC transition.

(see Fig. 4.6 (e) and (f)). The time dependent narrowing/widening of the DNA binding channel is described using the distance between centers of mass of the jaw-lobe and clamp modules, $d_{\rm JC}$ (see Fig. 4.6 (a)). The distance $d_{\rm JC} \approx 72\mathring{A}$ at $t = 0$, and decreases gradually for about $3\mathring{A}$ upon the F1→F2 transition at $8\mu s$. When the F2→EC transition is triggered at $t = 32\mu s$, $d_{\rm JC}$ decreases rapidly for about $7\mathring{A}$. In both transitions, the clamp module moves closer to the jaw-lobe module, and the width of the DNA binding channel is reduced. However, from the free form

to the EC form, the negatively charged DNA interacts with the positively charged active-site cleft, which results in the large channel narrowing of 7$\mathring{A}$. Similarly, the angle between $\overrightarrow{\mathrm{JS}}$ and $\overrightarrow{\mathrm{CS}}$, θ, is calculated to shown the dynamics of the channel (see Fig. 4.6 (d)). The gradual decrease of 3.5 degrees in the transition F1→F2, and the rapid decrease of 8 degrees in the transition F2→EC are observed, indicating that narrowing of the DNA channel is required in the EC form, and this narrowing process may be a results of the DNA binding.

As reference, the distance between centers of the jaw-lobe module and the shelf module, d_{JS}, and the distance between the centers of clamp module to the shelf module, d_{CS}, are also calculated (see Fig. 4.6 (b) and (c)). The distance d_{JS} changes less than 0.5$\mathring{A}$ in both transitions, meaning the jaw-lobe and shelf modules are relatively stationary, and the distance d_{CS} decreases about 3$\mathring{A}$ in the transition F1→F2, indicating the clamp moves closer to the shelf module and swings away from the back wall. Interestingly, in the transition F2→EC, d_{CS} decreases about 1.5$\mathring{A}$ and increases back by 1.5$\mathring{A}$, the relative distance in the EC is kept the same.

4.4.3 Formation of multiple native contacts between the clamp and shelf modules in the F1→F2 transition triggers the back-forth motion.

Several residues on Rpb1 are found to make contacts with the residues on Rpb5 to trigger the back-forth motion. In detail, the distance between Asp155A (A refers to Rpb1) and Gly122E (E refers to Rpb5) increase from 18$\mathring{A}$ in its native

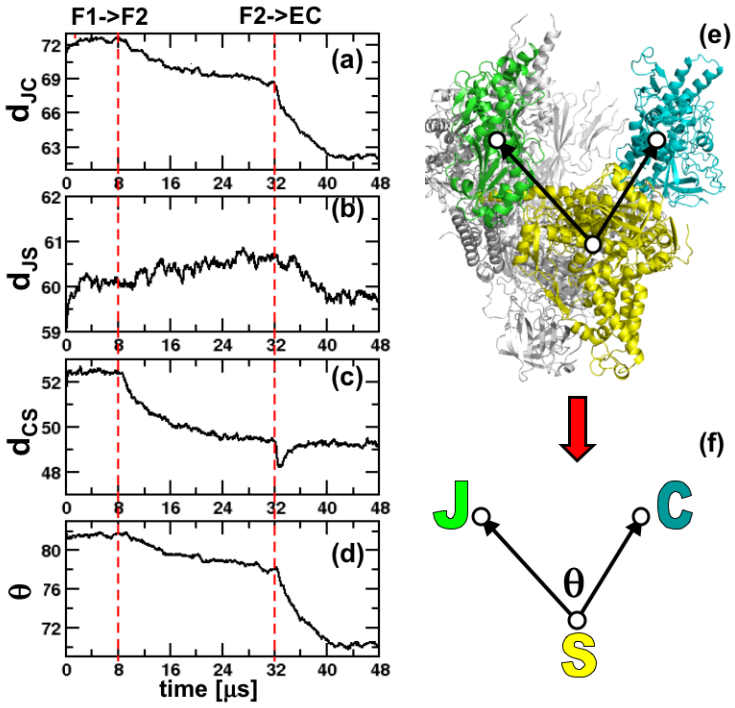


Figure 4.6: The global motion of clamp module, jaw-lobe module and shelf module. (a) d_{JC} as a function of time. (b) d_{JS} as a function of time. (c) d_{CS} as a function of time. (d) θ as a function of time. (e) Front view of Pol II, the clamp (C), jaw-lobe (J), and shelf (S) modules are colored cyan, green, and yellow, respectively. (f) The organization of clamp, jaw, and shelf modules.

structure in the F1 to about 20$\mathring{A}$ upon equilibration. During the F1→F2 transition, this distance drops slowly drop to 8$\mathring{A}$ (see Fig. 4.7 (a)). Another residue pair, Asp156A-Lys122E, which is essentially a salt-bridge interaction, is found to have similar distance changes in the F1→F2 transition. These large distance changes of about 12$\mathring{A}$ are assisted by formation and rupture of other contacts. For example, rupture of the contacts between residues Gly338A and Arg344A in switch 2, rupture of the contacts Ile336A-Met1152B (B refers to Rpb2) and Asn339A-Met1152B between the switch 2 and the switch 4. Formation of the contacts Asn1390A-Ile1408A and Ala1396A-Leu1409A in the switch 1.

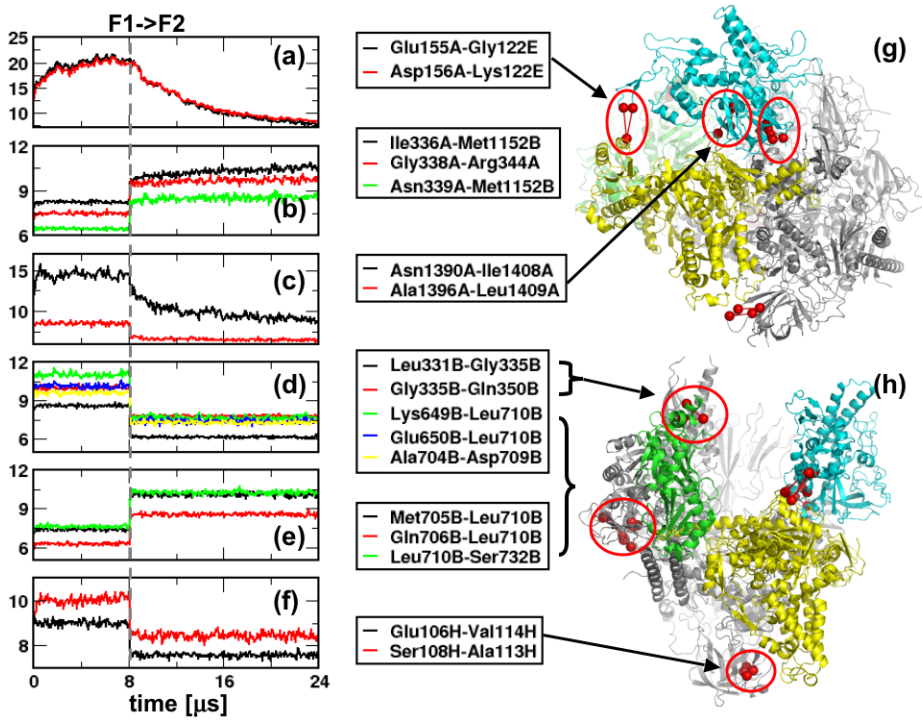


Figure 4.7: The evolution of distances of multiple contacts in F1→F2 transition. (a)-(f) Distance changes between contacted residues as a function of time. Residues involved in forming or breaking contacts are labeled on the right of each panel and shown on the structures with red spheres. (g)-(h) Structure of Pol II in F1 is used and side and front views are given.

Besides these contacts that are responsible for the back-forth motion of the clamp module, there are contacts formed/ruptured in the jaw-lobe module, such as Leu331B-Gly335B and Gly335B-Gln350B. In addition, there are contacts formed and broken in the core module near the jaw-lobe and shelf modules, such as Lys649B-Leu710B, Glu650B-Lue710B, Ala704B-Asp709B, Met705B-Leu710B, Gln706B-Leu710B, Leu710B-Ser732B, Glu106H-Val114H, and Ser108H-Ala113H. All the contacts that are not responsible for the back-forth motion are formed/ruptured relative fast (see Fig. 4.7 (d)-(f)) while the contacts formed in or between the switches take longer time to reach equilibrium (see Fig. 4.7 (a)-(c)).

4.4.4 Ordering of multiple switches in F2→EC transitions in the active center.

In the F2→EC transition, the dominant motion is the opening of the active-site cleft through the open-close and the outside-in motion of the clamp. In order to obtain the dynamics of the detailed conformational changes, we explicitly analyzed the distances changes of the contacted residue pairs. It is found that the DNA-RNA hybrid binding related transition emphasizes the ordering of several previously identified "switches" [129] which locates near the active center (see Fig. 4.4 and 4.8). In what follows, we will discuss the structural changes of the switches according to the time they occur in the F2→EC transition (see Fig. 4.8).

First, rupture of the salt-bridges between residues Lys330A and Glu333A in the switch 2 leads to unfolding of the helical turn in the switch 2, which causes Lys330A to move closer to the switch 1 and forms of the salt-bridge between with Glu1403A in the switch 1. (Fig. 4.8, (a) and (b)). Salt-bridges form in the switch 1 between Arg 1399A and Glu1403A, and Arg1399A and Glu1404A, which helps the switch 1 to fold to a helical structure (Fig. 4.8 (c)). On the other hand, salt-bridges form in the switch 4 between Arg1150B and Asp1153B, which results in the structural transition of the switch 4 from coil to turn (Fig. 4.8, (d)). The salt-bridge between Glu343A (switch 2) and Arg1156B (switch 4) is also formed, and the switch 2 interacts with the switch 4. At the same time, the switch 1 and the switch 4 become completely ordered upon the formation of the salt-bridges between Arg1386A and Glu1403A (switch 1), and between Lys1148B and Glu1153B (switch

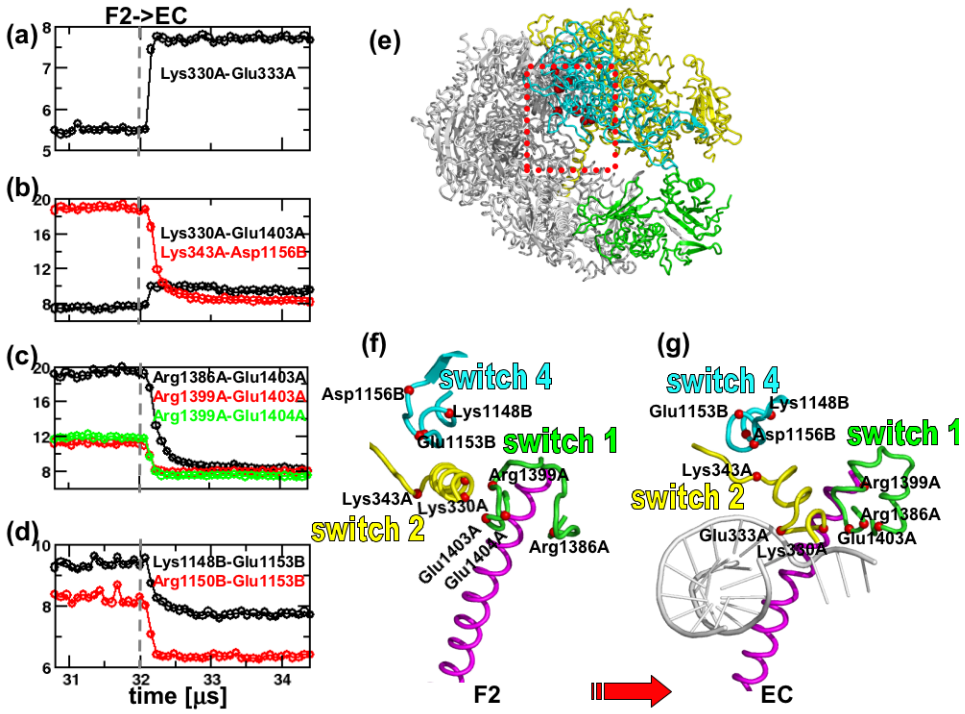


Figure 4.8: The evolution of distances of multiple salt-bridges in F2→EC transition. (a)-(d) Distance changes of the salt-bridges as a function of time. The four panels are placed from top to bottom according to the time sequence of these events. (e) Residues involved in forming or breaking of these salt-bridges are shown on the structure. (f)-(g) Multiple switches that are involved in F2→EC are shown in F2 and EC. Switch 1, switch 2 and switch 4 are shown in tube and colored according to secondary structure. Residues involved in salt-bridge formation or rupture are shown with spheres and labeled.

4) (Fig. 4.8 (a) and (d)). To conclude, the location of the switch 2 between the switch 1 and the switch 4 decides that the switch 2 mediates the ordering of these switches, and the interaction between them are largely electrostatic and related to the DNA-RNA hybrid binding.

4.5 Concluding Remarks

By applying the two methods, NMA and Brownian dynamics simulation to the complex Pol II system, we have obtained fundamental insights into the dynamics of Pol II transitions between various states. Besides the well-known open-close motion of the clamp, the back-forth and outside-in motions are discovered to be involved in the conformational changes of the Pol II. We have argued that although the requirement of the heterodimer Rpb4/7 in initiation and the fact that the heterodimer would restrict the open-close motion of the Pol II have ruled out the role of the open-close motion in initiation, the back-forth and outside-in motions could play very important roles in the initiation and elongation processes.

In addition, the finding that an interface interactions between Rpb1 and Rpb5 triggers the large scale back-forth motion in the F1→F2 indicate that it is possible to restrict these motions by adding disulfide bonds on the interface. This is also true for the F2→EC where multiple "switches" can be turned "on" or "off" to control the binding affinity of the hybrid bind sites. The sequence of the events observed in the simulations can also be tested by sequentially constraining the identified salt-bridges and measuring the changes in the global motion.

Bibliography

[1] Doi M, Edwards S (1986) *The theory of polymer dynamics* (Oxford University Press, New York), 1st edition.

[2] Thirumalai D, Hyeon C (2005) RNA and protein folding: Common themes and variations. *Biochemistry* 44:4957–4970.

[3] Boehr DD, McElheny D, Dyson HJ, Wright PE (2006) The dynamic energy landscape of dihydrofolate reductase catalysis. *Science* 313:1638–1642.

[4] Wolf-Watz M et al. (2004) Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nat. Struct. Mol. Biol.* 11:945–949.

[5] Boehr DD, Dyson HJ, Wright PE (2006) An NMR perspective on enzyme dynamics. *Chem. Rev.* 106:3055–3079.

[6] Hammes-Schiffer S, Benkovic SJ (2006) Relating protein motion to catalysis. *Annu. Rev. Biochem.* 75:519–541.

[7] Olsson MHM, Parson WW, Warshel A (2006) Dynamical contributions to enzyme catalysis: Critical tests of a popular hypothesis. *Chem. Rev.* 106:1737–1756.

[8] Henzler-Wildman KA et al. (2007) A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* 450:913–U927.

[9] Benkovic SJ, Hammes-Schiffer S (2003) A perspective on enzyme catalysis. *Science* 301:1196–1202.

[10] Hammes GG, Zhang ZQ, Rajagopalan PT (2003) Conformational changes in enzyme catalysis: Single-molecule and ensemble kinetics of dihydrofollate reductase. *Biochemistry* 42:8651–8652.

[11] Monod J, Wyman J, Changeux JP (1965) On nature of allosteric transitions - a plausible model. *J. Mol. Biol.* 12:88–118.

[12] Holt JM, Ackers GK (1995) The pathway of allosteric control as revealed by hemoglobin intermediate states. *FASEB J.* 9:210–218.

[13] Schnell JR, Dyson HJ, Wright PE (2004) Structure, dynamics, and catalytic function of dihydrofolate reductase. *Annu. Rev. Biophys. Biomol. Struct.* 33:119–140.

[14] Agarwal PK, Billeter SR, Rajagopalan PTR, Benkovic SJ, Hammes-Schiffer S (2002) Network of coupled promoting motions in enzyme catalysis. *Proc. Natl. Acad. Sci.* 99:2794–2799.

[15] Agarwal PK, Billeter SR, Hammes-Schiffer S (2002) Nuclear quantum effects and enzyme dynamics in dihydrofolate reductase catalysis. *J. Phys. Chem. B* 106:3283–3293.

[16] Wong K, Watney J, Hammes-Schiffer S (2004) Analysis of electrostatics and correlated motions for hydride transfer in dihydrofolate reductase. *J. Phys. Chem. B* 108:12231–12241.

[17] Wong KF, Selzer T, Benkovic SJ, Hammes-Schiffer S (2005) Impact of distal mutations on the network of coupled motions correlated to hydride transfer in dihydrofolate reductase. *Proc. Natl. Acad. Sci.* 102:6807–6812.

[18] Rajagopalan PT, Lutz S, Benkovic SJ (2002) Coupling interactions of distal residues enhance dihydrofolate reductase catalysis: mutational effects on hydride transfer rates. *Biochemistry* 41:12618–12628.

[19] Cannon WR, Singleton SF, Benkovic SJ (1996) A perspective on biological catalysis. *Nat. Struct. Biol.* 3:821–833.

[20] Radkiewicz JL, Brooks CL (2000) Protein dynamics in enzymatic catalysis: Exploration of dihydrofolate reductase. *J. Am. Chem. Soc.* 122:225–231.

[21] Rod TH, Radkiewicz JL, Brooks CL (2003) Correlated motion and the effect of distal mutations in dihydrofolate reductase. *Proc. Natl. Acad. Sci.* 100:6980–6985.

[22] Thorpe IF, Brooks CL (2003) Barriers to hydride transfer in wild type and mutant dihydrofolate reductase from E. coli. *J. Phys. Chem. B* 107:14042–14051.

[23] Thorpe IF, Brooks CL (2004) The coupling of structural fluctuations to hydride transfer in dihydrofolate reductase. *Proteins* 57:444–457.

[24] Garcia-Viloca M, Truhlar DG, Gao JL (2003) Reaction-path energetics and kinetics of the hydride transfer reaction catalyzed by dihydrofolate reductase. *Biochemistry* 42:13558–13575.

[25] Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286:295–299.

[26] Suel GM, Lockless SW, Wall MA, Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* 10:232–232.

[27] Lee J et al. (2008) Surface sites for engineering allosteric control in proteins. *Science* 322:438–442.

[28] Levitt M, Sander C, Stern PS (1985) Protein normal-mode dynamics - trypsin-inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.* 181:423–447.

[29] Bahar I, Rader AJ (2005) Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.* 15:586–592.

[30] Zheng WJ, Brooks BR, Doniach S, Thirumalai D (2005) Network of dynamically important residues in the open/closed transition in polymerases is strongly conserved. *Structure* 13:565–577.

[31] Zheng WJ, Brooks BR, Thirumalai D (2006) Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proc. Natl. Acad. Sci.* 103:7664–7669.

[32] Zheng W, Brooks BR, Thirumalai D (2007) Allosteric transitions in the chaperonin groel are captured by a dominant normal mode that is most robust to sequence variations. *Biophys. J.* 93:2289–2299.

[33] Tehver R, Chen J, Thirumalai D (2009) Allostery wiring diagrams in the transitions that drive the groel reaction cycle. *J. Mol. Biol.* 387:390–406.

[34] Hyeon C, Thirumalai D (2005) Mechanical unfolding of RNA hairpins. *Proc. Natl. Acad. Sci.* 102:6789–6794.

[35] Hyeon C, Dima RI, Thirumalai D (2006) Pathways and kinetic barriers in mechanical unfolding and refolding of RNA and proteins. *Structure* 14:1633–1645.

[36] Suh WC, Ross W, Record MT (1993) 2 open complexes and a requirement for mg^{2+} to open the lambda-p(r) transcription start site. *Science* 259:358–361.

[37] deHaseth PL, Zupancic ML, Record MT (1998) RNA polymerase-promoter interactions: the comings and goings of RNA polymerase. *J. Bacteriol.* 180:3019–3025.

[38] Sclavi B et al. (2005) Real-time characterization of intermediates in the pathway to open complex formation by Escherichia coli RNA polymerase at the T7A1 promoter. *Proc. Natl. Acad. Sci.* 102:4706–4711.

[39] Davis CA, Bingman CA, Landick R, Record MT, Saecker RM (2007) Real-time footprinting of DNA in the first kinetically significant intermediate in open complex formation by Escherichia coli RNA polymerase. *Proc. Natl. Acad. Sci.* 104:7833–7838.

[40] Mekler V et al. (2002) Structural organization of bacterial RNA polymerase holoenzyme and the RNA polymerase-promoter open complex. *Cell* 108:599–614.

[41] Murakami KS, Masuda S, Darst SA (2002) Structural basis of transcription initiation: RNA polymerase holoenzyme at 4 angstrom resolution. *Science* 296:1280–1284.

[42] Rees WA, Keller RW, Vesenka JP, Yang GL, Bustamante C (1993) Evidence of DNA bending in transcription complexes imaged by scanning force microscopy. *Science* 260:1646–1649.

[43] Rippe K, Guthold M, vonHippel PH, Bustamante C (1997) Transcriptional activation via DNA-looping: Visualization of intermediates in the activation pathway of E. coli RNA polymerase $\cdot\sigma^{54}$ holoenzyme by scanning force microscopy. *J. Mol. Biol.* 270:125–138.

[44] Rivetti C, Guthold M, Bustamante C (1999) Wrapping of DNA around the E.coli RNA polymerase open promoter complex. *EMBO J.* 18:4464–4475.

[45] Dima RI, Thirumalai D (2006) Determination of network of residues that regulate allostery in protein families using sequence analysis. *Protein Sci.* 15:258–268.

[46] Getz G, Levine E, Domany E (2000) Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci.* 97:12079–12084.

[47] Blatt M, Wiseman S, Domany E (1996) Superparamagnetic clustering of data. *Phys. Rev. Lett.* 76:3251–3254.

[48] Wiseman S, Blatt M, Domany E (1998) Superparamagnetic clustering of data. *Physical Review E* 57:3767–3783.

[49] Marques O, Sanejouand YH (1995) Hinge-bending motion in citrate synthase arising from normal mode calculations. *Proteins-Structure Function and Genetics* 23:557–560.

[50] Hyeon C, Thirumalai D (2007) Mechanical unfolding of RNA: From hairpins to structures with internal multiloops. *Biophys. J.* 92:731–743.

[51] Allen MP, Tildesley DJ (1987) *Computer simulation of liquids* (Clarendon Press, New York, NY, USA).

[52] Hyeon C, Lorimer GH, Thirumalai D (2006) Dynamics of allosteric transitions in GroEL. *Proc. Natl. Acad. Sci.* 103:18939–18944.

[53] Gardiner CW (2004) *Handbook of Stochastic Methods: for Physics, Chemistry and the Natural Sciences (Springer Series in Synergetics)* (Springer), 3rd edition.

[54] Chen J, Dima RI, Thirumalai D (2007) Allosteric communication in dihydrofolate reductase: Signaling network and pathways for closed to occluded transition and back. *J. Mol. Biol.* 374:250–266.

[55] Veitshans T, Klimov D, Thirumalai D (1997) Protein folding kinetics: Timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Folding & Design* 2:1–22.

[56] Polshakov VI, Birdsall B, Feeney J (1999) Characterization of rates of ring-flipping in trimethoprim in its ternary complexes with lactobacillus casei dihydrofolate reductase and coenzyme analogues. *Biochemistry* 38:15962–15969.

[57] Osborne MJ, Schnell J, Benkovic SJ, Dyson HJ, Wright PE (2001) Backbone dynamics in dihydrofolate reductase complexes: Role of loop flexibility in the catalytic mechanism. *Biochemistry* 40:9846–9859.

[58] McElheny D, Schnell JR, Lansing JC, Dyson HJ, Wright PE (2005) Defining the role of active-site loop fluctuations in dihydrofolate reductase catalysis. *Proc. Natl. Acad. Sci.* 102:5032–5037.

[59] Cameron CE, Benkovic SJ (1997) Evidence for a functional role of the dynamics of glycine-121 of Escherichia coli dihydrofolate reductase obtained from kinetic analysis of a site-directed mutant. *Biochemistry* 36:15792–15800.

[60] Wang L, Goodey NM, Benkovic SJ, Kohen A (2006) Coordinated effects of distal mutations on environmentally coupled tunneling in dihydrofolate reductase. *Proc. Natl. Acad. Sci.* 103:15753–15758.

[61] Hammes GG (1964) Mechanism of enzyme catalysis. *Nature* 204:342–343.

[62] Miller GP, Benkovic SJ (1998) Stretching exercises–flexibility in dihydrofolate reductase catalysis. *Chem. Biol.* 5:R105–113.

[63] Berg J, Tymoczko JT, Stryer L (2002) *Biochemistry* (New York: W. H. Freeman and Co.), Fifth edition.

[64] Matthews DA et al. (1977) Dihydrofolate-reductase - X-Ray structure of binary complex with methotrexate. *Science* 197:452–455.

[65] Sawaya MR, Kraut J (1997) Loop and subdomain movements in the mechanism of Escherichia coli dihydrofolate reductase: Crystallographic evidence. *Biochemistry* 36:586–603.

[66] Venkitakrishnan RP et al. (2004) Conformational changes in the active site loops of dihydrofolate reductase during the catalytic cycle. *Biochemistry* 43:16046–16055.

[67] Changeux JP, Edelstein SJ (2005) Allosteric mechanisms of signal transduction. *Science* 308:1424–1428.

[68] Thirumalai D, Lorimer GH (2001) Chaperonin-mediated protein folding. *Annu. Rev. Biophys. Biomol. Struct.* 30:245–269.

[69] Hatley ME, Lockless SW, Gibson SK, Gilman AG, Ranganathan R (2003) Allosteric determinants in guanine nucleotide-binding proteins. *Proc. Natl. Acad. Sci.* 100:14445–14450.

[70] Jain RK, Ranganathan R (2004) Local complexity of amino acid interactions in a protein core. *Proc. Natl. Acad. Sci.* 101:111–116.

[71] Shulman AI, Larson C, Mangelsdorf DJ, Ranganathan R (2004) Structural determinants of allosteric ligand activation in RXR heterodimers. *Cell* 116:417–429.

[72] Tang C, Iwahara J, Clore GM (2006) Visualization of transient encounter complexes in protein-protein association. *Nature* 444:383–386.

[73] Bateman A et al. (2002) The Pfam protein families database. *Nucleic Acids Res.* 30:276–280.

[74] Thompson JD, Higgins DG, Gibson TJ (1994) Clustal-w - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673–4680.

[75] Domany E (1999) Superparamagnetic clustering of data - the definitive solution of an ill-posed problem. *Physica A* 263:158–169.

[76] Stan G, Thirumalai D, Lorimer GH, Brooks BR (2003) Annealing function of groel: structural and bioinformatic analysis. *Biophys. Chem.* 100:453–467.

[77] Howell E, Villafranca J, Warren M, Oatley S, J. K (1986) Functional role of aspartic acid-27 in dihydrofolate reductase revealed by mutagenesis. *Science* 231:1123–1128.

[78] Wang L, Goodey NM, Benkovic SJ, Kohen A (2006) The role of enzyme dynamics and tunnelling in catalysing hydride transfer: studies of distal mutants of dihydrofolate reductase. *Philos Trans R Soc Lond B Biol Sci* 361:1307–1315.

[79] Mauldin RV, Carroll MJ, Lee AL (2009) Dynamic dysfunction in dihydrofolate reductase results from antifolate drug binding: Modulation of dynamics within a structural state. *Structure* 17:386 – 394.

[80] Liu HB, Warshel A (2007) The catalytic effect of dihydrofolate reductase and its mutants is determined by reorganization energies. *Biochemistry* 46:6011–6025.

[81] Sergi A, Watney JB, Wong KF, Hammes-Schiffer S (2006) Freezing a single distal motion in dihydrofolate reductase. *J. Phys. Chem. B* 110:2435–2441.

[82] Tsai CJ, Kumar S, Ma BY, Nussinov R (1999) Folding funnels, binding funnels, and protein function. *Protein Sci.* 8:1181–1190.

[83] Koshland DE (1958) Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci.* 44:98–104.

[84] Swanwick RS, Maglia G, Tey LH, Allemann RK (2006) Coupling of protein motions and hydrogen transfer during catalysis by Escherichia coli dihydrofolate reductase. *Biochem. J.* 394:259–265.

[85] Borukhov S, Nudler E (2003) RNA polymerase holoenzyme: structure, function and biological implications. *Curr. Opin. Microbiol.* 6:93–100.

[86] Darst SA, Kubalek EW, Kornberg RD (1989) 3-Dimensional structure of Escherichia-Coli RNA-polymerase holoenzyme determined by electron crystallography. *Nature* 340:730–732.

[87] Zhang GY et al. (1999) Crystal structure of Thermus aquaticus core RNA polymerase at 3.3 Angstrom resolution. *Cell* 98:811–824.

[88] Finn RD, Orlova EV, Gowen B, Buck M, van Heel M (2000) Escherichia coli RNA polymerase core and holoenzyme structures. *EMBO J.* 19:6833–6844.

[89] Darst SA (2001) Bacterial RNA polymerase. *Curr. Opin. Struct. Biol.* 11:155–162.

[90] Cramer P (2002) Multisubunit RNA polymerases. *Curr. Opin. Struct. Biol.* 12:89–97.

[91] Murakami KS, Masuda S, Campbell EA, Muzzin O, Darst SA (2002) Structural basis of transcription initiation: An RNA polymerase holoenzyme-DNA complex. *Science* 296:1285–1290.

[92] Vassylyev DG et al. (2002) Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 Angstrom resolution. *Nature* 417:712–719.

[93] Craig ML et al. (1998) DNA footprints of the two kinetically significant intermediates in formation of an RNA polymerase-promoter open complex: Evidence that interactions with start site and downstream DNA induce sequential conformational changes in polymerase and DNA. *J. Mol. Biol.* 283:741–756.

[94] Chen YF, Helmann JD (1997) DNA-melting at the bacillus subtilis flagellin promoter nucleates near -10 and expands unidirectionally. *J. Mol. Biol.* 267:47–59.

[95] Kapanidis AN et al. (2006) Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism. *Science* 314:1144–1147.

[96] Revyakin A, Liu CY, Ebright RH, Strick TR (2006) Abortive initiation and productive initiation by RNA polymerase involve DNA scrunching. *Science* 314:1139–1143.

[97] Murakami KS, Darst SA (2003) Bacterial RNA polymerases: the wholo story. *Curr. Opin. Struct. Biol.* 13:31–39.

[98] Gralla JD (2000) Signaling through sigma. *Nat. Struct. Biol.* 7:530–532.

[99] Record MT, Jr, Reznikoff WS, Craig ML, McQuade KL, Schlax PJ (1996) *Escherichia coli RNA polymerase ($E\sigma^{70}$), promoters, and the kinetics of the steps of transcription onitiation.*, ed Neidhardt FC (ASM Press, Washington, D.C.), 2nd edition, pp 792–820.

[100] Bai L, Santangelo TJ, Wang MD (2006) Single-molecule analysis of RNA polymerase transcription. *Annu. Rev. Biophys. Biomol. Struct.* 35:343–360.

[101] Burgess RR, Travers AA, Dunn JJ, Bautz EKF (1969) Factor stimulating transcription by RNA polymerase. *Nature* 221:43–46.

[102] Campbell EA et al. (2002) Structure of the bacterial RNA polymerase promoter specificity sigma subunit. *Mol. Cell* 9:527–539.

[103] Gross CA et al. (1998) The functional and regulatory roles of sigma factors in transcription. *Cold Spring Harbor Symp. Quant. Biol.* 63:141–155.

[104] Tropp BE (2007) *Molecular Biology Genes to Proteins* (Jones and Bartlett publishers), 3rd edition.

[105] Aiyar SE, Juang YL, Helmann JD, deHaseth PL (1994) Mutations in sigma factor that affect the temperature dependence of transcription from a promoter, but not from a mismatch bubble in double-stranded DNA. *Biochemistry* 33:11501–11506.

[106] Juang YL, Helmann JD (1994) A promoter melting region in the primary sigma-factor of bacillus-subtilis - identification of functionally important aromatic-amino-acids. *J. Mol. Biol.* 235:1470–1488.

[107] Rong JC, Helmann JD (1994) Genetic and physiological-studies of bacillus-subtilis sigma(a) mutants defective in promoter melting. *J. Bacteriol.* 176:5218–5224.

[108] Tang GQ, Patel SS (2006) T7 RNA polymerase-induced bending of promoter DNA is coupled to DNA opening. *Biochemistry* 45:4936–4946.

[109] Tang GQ, Patel SS (2006) Rapid binding of T7 RNA polymerase is followed by simultaneous bending and opening of the promoter DNA. *Biochemistry* 45:4947–4956.

[110] Tang GQ, Roy R, Ha T, Patel SS (2008) Transcription initiation in a single-subunit RNA polymerase proceeds through DNA scrunching and rotation of the n-terminal subdornains. *Mol. Cell* 30:567–577.

[111] Severinov K, Darst SA (1997) A mutant RNA polymerase that forms unusual open promoter complexes. *Proc. Natl. Acad. Sci.* 94:13481–13486.

[112] Rogozina A, Zaychikov E, Buckle M, Heumann H, Sclavi B (2009) DNA melting by RNA polymerase at the T7A1 promoter precedes the rate-limiting step at 37 degrees C and results in the accumulation of an off-pathway intermediate. *Nuc. Acid. Res.* 37:5390–5404.

[113] Schroeder LA et al. (2009) Evidence for a tyrosine-adenine stacking interaction and for a short-lived open intermediate subsequent to initial binding of Escherichia coli RNA polymerase to promoter DNA. *J. Mol. Biol.* 385:339–349.

[114] Vuthoori S, Bowers CW, McCracken A, Dombroski AJ, Hinton DM (2001) Domain 1.1 of the sigma(70) subunit of escherichia coli RNA polymerase modulates the formation of stable polymerase/promoter complexes. *J. Mol. Biol.* 309:561–572.

[115] Koga N, Takada S (2006) Folding-based molecular simulations reveal mechanisms of the rotary motor F-1-ATPase. *Proc. Natl. Acad. Sci.* 103:5367–5372.

[116] Thirumalai D, Klimov DK, Woodson SA (1997) Kinetic partitioning mechanism as a unifying theme in the folding of biomolecules. *Theor. Chem. Acc.* 96:14–22.

[117] Thirumalai D, Ha BY (1997) Statistical mechanics of semiflexible chains: A meanfield variational approach. *arXiv:cond-mat* 9705200.

[118] Hyeon C, Thirumalai D (2008) Multiple probes are required to explore and control the rugged energy landscape of RNA hairpins. *J. Am. Chem. Soc.* 130:1538–1539.

[119] Roeder RG (1996) The role of general initiation factors in transcription by RNA polymerase II. *Trends in Biochemical Sciences* 21:327–335.

[120] Kornberg RD (1999) Eukaryotic transcriptional control (reprinted from trends in biochemical science, vol 12, dec., 1999). *Trends in Cell Biology* 9:M46–M49.

[121] Lee TI, Young RA (2000) Transcription of eukaryotic protein-coding genes. *Annual Review of Genetics* 34:77–137.

[122] Conaway JW, Conaway RC (1999) Transcription elongation and human disease. *Annual Review of Biochemistry* 68:301–319.

[123] Shilatifard A (1998) Factors regulating the transcriptional elongation activity of RNA polymerase II. *Faseb Journal* 12:1437–1446.

[124] Brueckner F, Ortiz J, Cramer P (2009) A movie of the RNA polymerase nucleotide addition cycle. *Current Opinion in Structural Biology* 19:294 – 299.

[125] Vassylyev DG (2009) Elongation by RNA polymerase: a race through roadblocks. *Curr. Opin. Struct. Biol.* 19:691–700.

[126] Hirose Y, Manley JL (2000) RNA polymerase II and the integration of nuclear events. *Genes & Development* 14:1415–1429.

[127] Proudfoot N (2000) Connecting transcription to messenger RNA processing. *Trends in Biochemical Sciences* 25:290–293.

[128] Cramer P, Bushnell DA, Kornberg RD (2001) Structural basis of transcription: RNA polymerase II at 2.8 Angstrom ngstrom resolution. *Science* 292:1863–1876.

[129] Gnatt AL, Cramer P, Fu JH, Bushnell DA, Kornberg RD (2001) Structural basis of transcription: An RNA polymerase II elongation complex at 3.3 angstrom resolution. *Science* 292:1876–1882.

[130] Cramer P, Arnold E (2009) Proteins: how RNA polymerases work. *Current Opinion in Structural Biology* 19:680 – 682.

[131] Chen J, Darst SA, Thirumalai D (2010) Transcription initiation triggered by bacterial RNA polymerase occurs in three steps. *Proc. Natl. Acad. Sci. U. S. A.* Submitted, March.

[132] Cramer P (2006) Deciphering the RNA polymerase II structure: a personal perspective. *Nat. Struct. Mol. Biol.* 13:1042–1044.

[133] Fu JH et al. (1999) Yeast RNA polymerase II at 5 Angstrom resolution. *Cell* 98:799–810.

[134] Bushnell DA, Kornberg RD (2003) Complete, 12-subunit RNA polymerase II at 4.1-A resolution: implications for the initiation of transcription. *Proc. Natl. Acad. Sci.* 100:6969–6973.

[135] Armache KJ, Kettenberger H, Cramer P (2003) Architecture of initiation-competent 12-subunit RNA polymerase II. *Proceedings of the National Academy of Sciences of the United States of America* 100:6964–6968.

[136] Cramer P (2004) RNA polymerase II structure: from core to functional complexes. *Curr. Opin. Genet. Dev.* 14:218–226.

[137] Mukhopadhyay J et al. (2008) The RNA polymerase "switch region" is a target for inhibitors. *Cell* 135:295 – 307.